Data reduction methods to study cancer susceptibility

Santaolalla Revenga, Aida

Awarding institution:
King's College London

Data reduction methods to study cancer susceptibility

Thesis presented in accordance with the requirements for the degree of Doctor of Philosophy

by
Aida Santaolalla Revenga

Translational Oncology and Urology Research (TOUR)
School of Cancer and Pharmaceutical Sciences
King's College London
United Kingdom

February 2018

Acknowledgement

When I started my journey in science, I was deeply intrigued about biology and its systemic level functioning and mathematics. This enthusiasm has driven years working in health sciences, first in bioinformatics, then health informatics and computational biology and finally in cancer epidemiology and biostatistics. This adventure somehow guided me into cancer epidemiology and biostatistics, disciplines that I am truly passionate about. In 2011 I was fortunate to join Prof Lars Holmberg's KCL team and afterwards Dr Mieke Van Hemelrijck's KCL team, where I immersed myself into a dynamic, translational and challenging environment where I learned about cancer, epidemiology and biostatistics. Dr Mieke Van Hemelrijck with her passion and her amazing work ethos encouraged me to pursue a PhD in Cancer Epidemiology, while working as a data manager for the group. She inspired me to continue learning and accept new challenges over the years that we have been working together. Therefore, I would like to thank her for her wonderful support and dedication during all these years.

I would also like to thank my second supervisor Dr Anita Grigoriadis for all her help and guidance with the Thesis projects and my third supervisor Prof Ton Coolen for his great support with the mathematical methods and the Terahertz project.

I would also like to thank Prof Lars Holmberg, a true role model for a data scientist through his global vision and his critical approach to science. I also have to say a big thank you to the all the members of the TOUR group for their help, their support and their dynamic, joyful and inspiring work – in particular Dr Hans Garmo for his advice with the statistical methods.

I also want to acknowledge the Swedish collaborators, the AMORIS team (Ingmar Jungner, Niklas Hammar, and Göran Walldius) for giving me the opportunity to work with their amazing database and for their guidance on the projects.

My friends and family deserve a special thank you. They have been always there supporting and comforting me during this amazing journey and specially in this last period where a second journey has started with our first child coming in April. My friends at KCL, Fara, Virginia, Erika and Salpie, for the nice lunches and laughs and my friends in Spain, Pepe, Sara, Berta, Kris, Olga, Adri and Patri, who always stayed in touch. My parents, Carmen and Angel, have always encouraged me to follow my instinct and pursue my dreams, and showed me that perseverance and curiosity are driving forces in live. My sister Bea, my niece Mariana and Rui for their love and support during this adventure and my dog Mipa who kept me company every evening whilst writing this thesis. My grandmother, Lucinia, who is looking out for me every day.

My beloved Manuel deserves a very special acknowledgment. Without his help, and especially in this last period whilst also being pregnant, this thesis could not have been possible. He guides me through life, thanks for always be there with a smile.

Abstract

Background and Aims

Cancer burden continues to increase in an aging population and hence cancer data has evolved into complex and multidimensional datasets with the advent of the OMICs sciences. Pathogenesis varies between patients and presents an intricate gene-environment interplay, which is reflected by the multifactorial character of the population susceptibility to cancer.

The current thesis, therefore, aims to comprehend population susceptibility to cancer and heterogeneity of the disease by investigating new statistical approaches using multidimensional cancer datasets to ultimately develop effective stratification models for cancer risk, with the potential to improve cancer prevention and early detection.

Methods

The thesis is divided into two main areas of study: Population susceptibility to disease and Individuals' susceptibility to disease.

1. **Population susceptibility to disease.**

The following projects utilised data from the Apolipoprotein MOrtality RISk (AMORIS) study:

a. The Blood exposome

A subset of the internal-external blood exposome components were evaluated by exploring the reciprocity of 21 standard serum markers and 4 external factors following a four-step statistical analysis: correlation analysis, hierarchical clustering, principal component analysis and multivariable analysis of the variance (n=154,207).

b. Metabolic profiles to assess cancer risk and mortality

To identify metabolic profiles linked to carcinogenesis and mortality and their intrinsic associations, latent class analyses followed by multivariate Cox regression analyses were performed to characterise subgroups of individuals based on 19

standard blood biomarker measurements to reflect population heterogeneity (n=13,615).

2. Individuals' susceptibility to disease.

c. Discrimination of breast cancer tissue

Imaging data generated by scanning 44 ex vivo breast tissue samples, utilising a terahertz probe (n=257), was evaluated using a two-step statistical approach (Gaussian deconvolution processing followed by a Naïve Bayes Classifier) to distinguish malignant and benign breast tissue, with the ultimate aim to identify malignant tissue intraoperatively ensuring clear negative tumour margins in breast-conserving surgery.

Results

a) The subset of the blood exposome analysis in AMORIS showed a tight interaction between internal markers of related pathways such as iron markers, whilst less well-known correlations also appeared (Albumin and Calcium). External markers showed that males and lower education were associated with serum biomarker levels that might be indicative of worse health outcomes. The variability of the data was distributed among all the markers studied.

b) The metabolic profiles analysis in AMORIS identified four LCA metabolic profiles within the population: (1) normal values for all markers (63% of population); (2) abnormal values for lipids (22%); (3) abnormal values for liver functioning (9%); (4) abnormal values for iron and inflammation metabolism (6%). All metabolic profiles (classes 2-4) increased risk of cancer and mortality, compared to class 1 (e.g. HR for overall death was 1.26 (95%CI: 1.16 - 1.37), 1.67 (95%CI: 1.47 - 1.90), and 1.21 (95%CI: 1.05 - 1.41) for class 2, 3, and 4, respectively).

c) The Bayesian classifier for tumour tissue discrimination performed using the combined Gaussian derivatives, obtained the following values: 69%, 89%, 53%, 60%, 86%, for accuracy, sensitivity, specificity, positive predictive value and negative predictive value respectively. Tumour tissue was classified

correctly in more than 89% of the cases with an accuracy of 0.7 and sensitivity of 0.9.

Conclusion

The subset of the blood exposome studied presented a complex synergy between the internal-external components, which demonstrates the need of systemic approaches involving multiple markers capable of evaluating the internal biological and external environment when assessing health outcomes. Moreover, the LCA analysis indicated that internal blood markers, when assembled into meaningful metabolic profiles by optimised statistical methods, could help stratify the population for cancer risk and mortality and provide insight in cancer susceptibility and aetiology. Finally, the Bayesian classifier effectively discriminated malignant from benign breast tissue using TPI imaging data, however presenting moderate specificity, which suggests the potential clinical applicability of this method to improve the adequate excision of the margins in BCS surgery, if the specificity can be optimised.

Overall, the projects in this thesis demonstrate the capability of data reduction methods to explore cancer susceptibility and develop potentially effective stratification models, and highlight the importance of data driven approaches in the assessment of multifactorial diseases such as cancer when supported by robust statistical analysis.

Table of contents

List of tables

List of figures

List of abbreviations

AIC	Akaike Information Criterion
ALP	Alkaline Phosphatase
ALT	Alanine Amino Transferase
AMORIS	Apolipoprotein MOrtality RISk Study
ANOVA	Analysis of the variance
ApoA-I	Apolipoprotein A-1
ApoB	Apolipoprotein B
ATP	Adenosine Triphosphate
AST	Aspartate Amino Transferase
BC	Breast Cancer
BCS	Breast-Conserving Surgery
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CALAB	Central Automation Laboratory
CCI	Charlson Comorbidity Index
CRP	C-Reactive Protein
CVD	Cardiovascular Disease
DCIS	Ductal Carcinoma In Situ
DNA	Deoxyribonucleic Acid
EC	Endometrial Cancer
EM	Expectation-Maximization
ER	Estrogen Receptor
FA	Factor Analysis
FAMN	Fructosamine
FDR	False Discovery Rate
FE	Iron
Fe-PCT	Transferrin Saturations
FWER	Family-Wise Error Rate
GBD	Global Burden of Disease methodology
GGT	Gamma-glutamyl Transferase
HDL	High Density Lipoprotein
HER2	Human Epidermal Growth Factor Receptor 2
HR	Hazard Ratio
ICD-9	International Classification of Diseases 9th Revision
ICD-O/2	International Classification of Diseases for Oncology 2nd Revision
ICD-O/3	International Classification of Diseases for Oncology 3rd Revision
ICH	Immunohistochemistry
IDC	Invasive Ductal Carcinoma
ILC	Invasive Lobular Carcinoma
IMA	Intraoperative Margin Assessment
KHP	King's Health Partners
LCA	Latent Class Cluster Analysis

LDH	Lactate Dehydrogenase
LDL	Low Density Lipoprotein
MAP	Maximum A Posteriori Probability protocol
NADH	Reduced Nicotinamide Adenine Dinucleotide
NPV	Negative Predictive Value
NST	No Specific Type carcinoma
OR	Odds Ratio
PCA	Principal Component Analysis
PCa	Prostate Cancer
PLS	Partial Least Square
PPV	Positive Predictive Value
PR	Progesterone Receptor
PSA	Prostate Specific Antigen
PTKs	Receptor Protein-Tyrosine Kinases
RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
RR	Risk Ratio
SAS	Statistical Analysis Systems
SES	Socio-Economic Status
SNP	Single Nucleotide Polymorphism
TC	Total Cholesterol
TG	Triglycerides
THz	Terahertz
TIBC	Total Iron Binding Capacity
TPI	Terahertz Pulsed Imaging
URAT	Uric Acid
WBC	Leukocytes
WHO	World Health Organization
95%CI	95% Confidence Interval

Chapter I. Introduction and research objectives

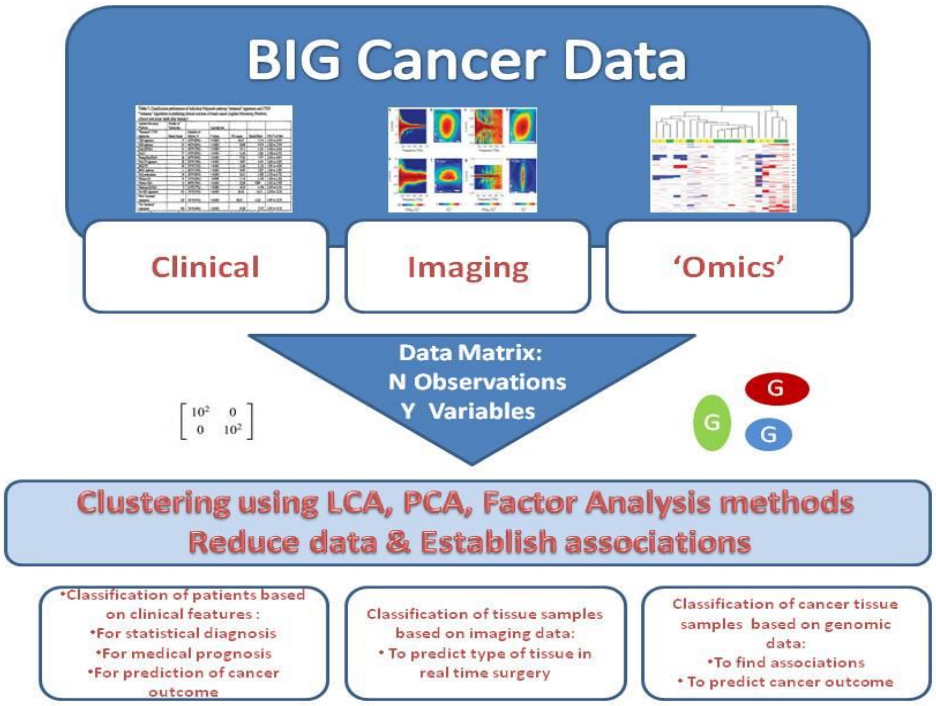
Cancer development involves multiple biological pathways and includes diverse genetic, molecular and clinical events, with pathogenesis varying between patients (1, 2). It is critical to identify which biomarkers are linked to carcinogenesis, as well as their associations, to get a better understanding of the complexity of cancer, specifically its causes and evolution (3, 4). This could improve cancer assessment and lead to early clinical interventions, better treatment options and improved patient outcomes (5).

Cancer data has been generated by the latest technologies, resulting in highly complex, shared multiple types of data and huge volumes of information. For instance, advances in 'omics' technologies have created many candidate markers with potential clinical value for cancer diagnosis and prognosis, but so far relatively few have made it into clinical practice (6, 7). Thus, more systematic-based approaches are needed to replace single biomarker analysis with multiple testing (multiple biomarkers) in large datasets (5, 8, 9). This will enhance the shift from a classic hypothesis driven (exposure to outcome) to a data driven approach (multiple exposures to an outcome), accelerating the discovery of biological mechanism links with disease and with population susceptibility to disease (10).

Several different data reductions methods have been used to explore the relation between biomarkers and disease in health sciences (11). Data reduction methods are defined as mathematical algorithms that decrease data dimensionality to help its exploration. Clustering is one such data exploratory tool that has been used in many areas such as clinical information, genomics and proteomics (12). The main idea is to group objects into meaningful classes that will describe the data. Related samples are clustered according to similarity coefficients. Another technique less used in cancer studies is Latent Class Analysis (LCA), a classification method to explore the underlying relationships between biomarkers, which groups subjects based on their latent classes instead of grouping biomarkers (13). Hence, these

analysis techniques reduce data to a manageable size allowing easier identification of associations, as described in Figure 1. The figure illustrates the concept behind the statistical approach used in this thesis.

Figure 1. Schematic representation of the concept followed in the statistical methodology. The methodology can be applied in different biomedical settings: clinical, imaging or genomics. A big cancer dataset contains a large amount of cancer data structured in a data matrix with multiple observations and variables that define these observations. The analysis aims to reduce the dimension of the data by finding subgroups in the datasets and establishing associations within the data.



With the overall goal of investigating statistical methods to stratify individuals based on their underlying risk of cancer whilst using multiple biomarkers, this thesis aimed to explore cancer susceptibility using data reduction methods in different biomedical settings. More specifically, this thesis comprises of the following two approaches to explore cancer susceptibility:

1. To investigate population susceptibility to cancer in a clinical setting, I explored whether data driven approaches could develop effective cancer risk stratification using the Swedish Apolipoprotein MOrtality RISk Cohort:

a. Project A: I investigated a subset of the blood exposome by (1) evaluating interactions in routinely assessed health biomarkers and (2) assessing how the external environment influences these interactions.
b. Project B: I evaluated how metabolic profiles can assess risk of cancer and mortality.

2. To investigate individual susceptibility to cancer in an imaging setting, I explored individuals' heterogeneity in an imaging diagnostic setting to improve breast cancer tissue discrimination intraoperatively, using material from King's Health Partners' Breast Cancer Biobank:
    a. Project C: I evaluated how imaging data produced by a terahertz probe can discriminate between different breast tissue samples, with the ultimate aim of using this probe intraoperatively in breast-conserving surgery to predict positive margins.

Overall, this thesis aimed to enhance our understanding of the use of data reduction methods in different biomedical settings and comprehend how to implement these methods to assess cancer susceptibility more efficiently to ultimately develop effective cancer risk stratification tools.

The next chapter provides an overview of the current state of the art with respect to cancer epidemiology and current approaches to cancer risk assessment. Chapter III explores methods to reduce data dimensionality and optimize prediction – with a specific focus on the methods used in the three projects described above. The methods and results of the studies in the Swedish AMORIS database are described in Chapter IV, whereas Chapter V describes the findings of the imaging project. Finally, chapter VI is the concluding chapter, which interprets the results and provides guidance for future research.

Chapter II. Background on cancer heterogeneity

I. Cancer Epidemiology

a. Global burden of cancer

Cancer continues to be one of the major causes of death worldwide. The burden of cancer is increasing due to the aging of a growing population and the adoption of health risk behaviours, mainly in the developed countries (14).

In 2015, 17.5 million cancer cases and 8.7 million cancer deaths were estimated worldwide using the Global Burden of Disease methodology (GBS), establishing cancer as the second leading cause of death, following cardiovascular disease (CVD) (15).

Since 2005, incident cancer cases have increased by 33% and this number is expected to continue raising over the following years, with breast cancer (BC) being the most common cause of cancer accounting for around 2.4 million cases in 2015 (15). During the same period, cancer deaths have decreased in many countries, however the opposite trend is seen some countries, such as in the Sub-Saharan Africa region (Figure 2). Cancer survival tend to be worse in developing countries, mainly because of the lack of screening policies and the limited access to treatments (16).

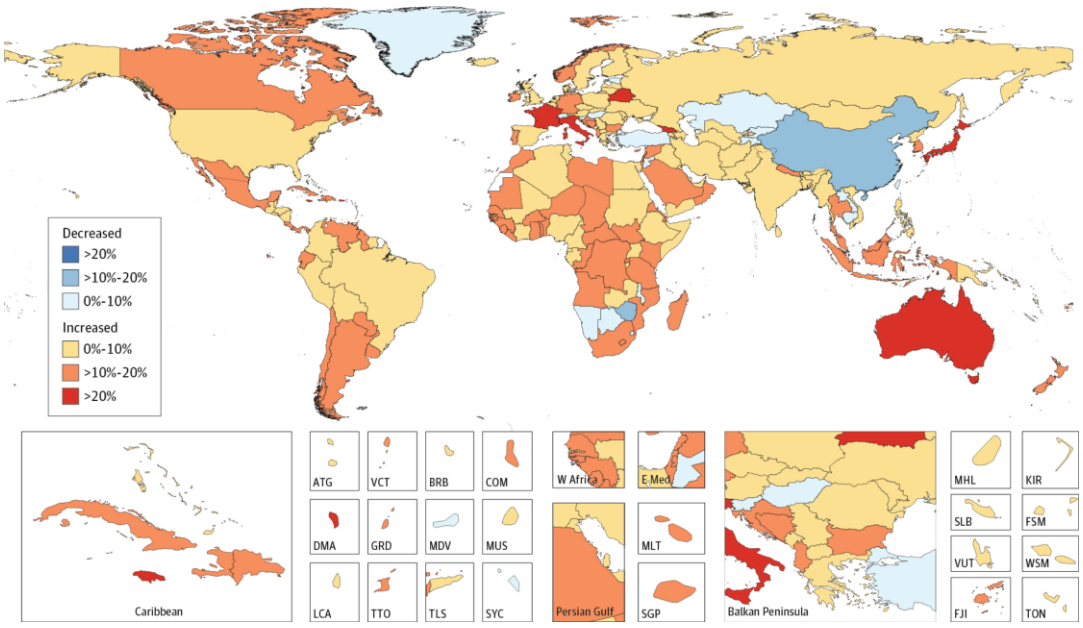
In men, prostate cancer (PCa) is the most common cancer worldwide with tracheal, bronchus, and lung cancer being the main causes of cancer death, while in women breast cancer remains the leading cause of cancer and cancer death (Figure 3) (15).

To meet the cancer challenge, personalized medicine has explored novel treatment approaches, such as T-cell engineering in Immunotherapy (17), improving life expectancy in patients in high income countries. In contrast, low income countries still face the need for early detection programs and accessible standard treatment

options for the population (14). Cancer prevention, early diagnosis and adequate treatments are thus key to fight cancer.

Figure 2. Relative Changes in Age-Standardized Cancer Incidence Rates (A) and Mortality Rates (B) in Both Sexes for All Cancers in 195 Countries or Territories from 2005 to 2015. Figure taken from the Global Burden of Disease Cancer, 2017 (15).

A)



B)

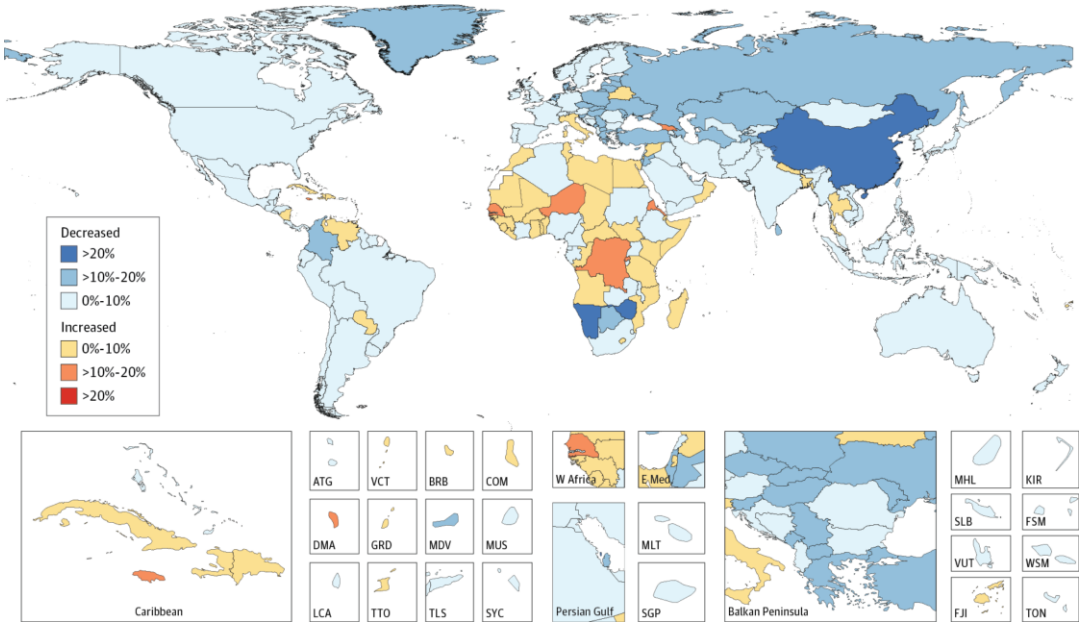
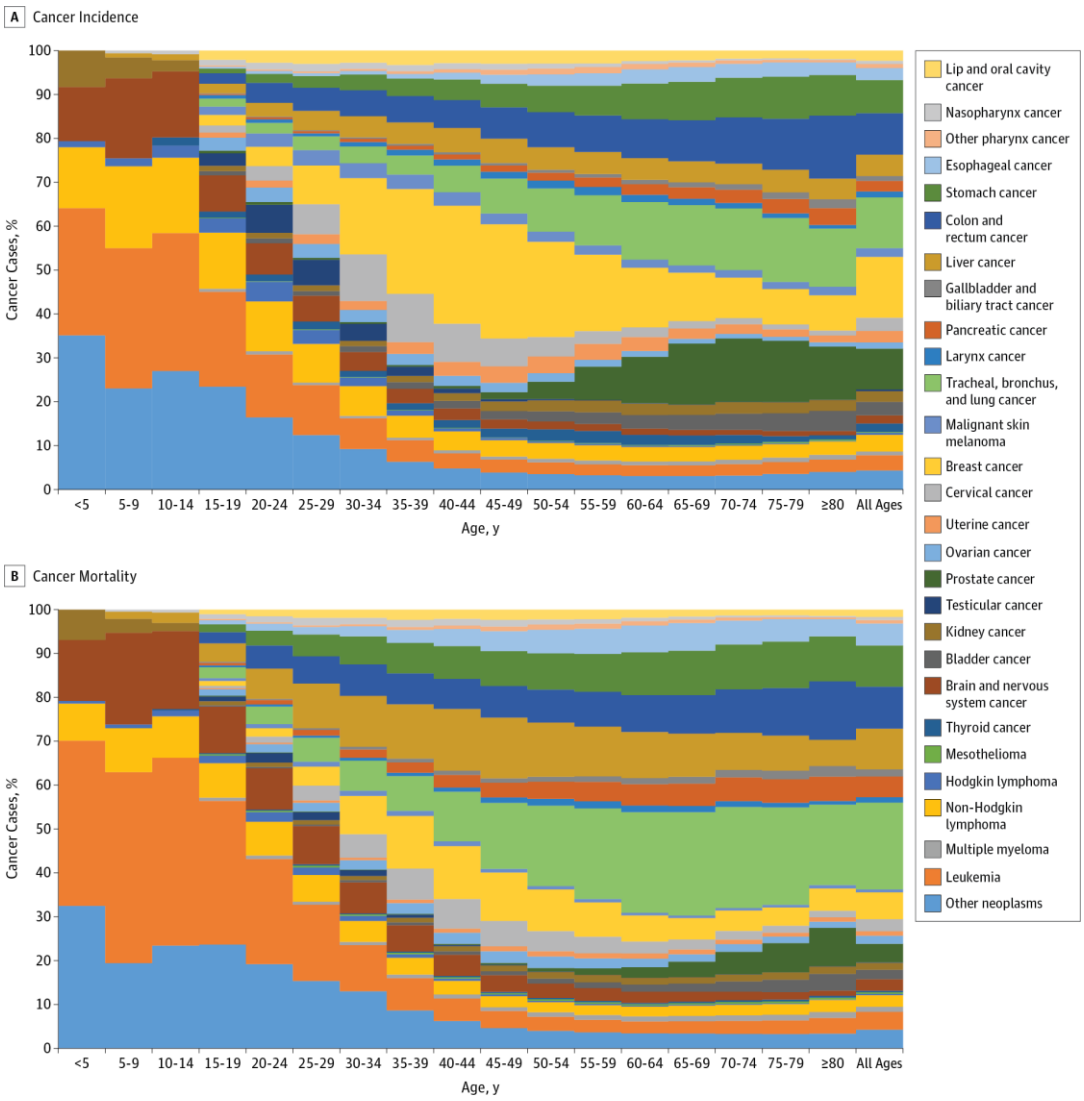


Figure 3. Age-Specific Global Contributions of Cancer Types to Total Cancer Incidence and Mortality for Both Sexes. Figure taken from the Global Burden of Disease Cancer, 2017 (15).



b. Carcinogenesis and hallmarks of cancer

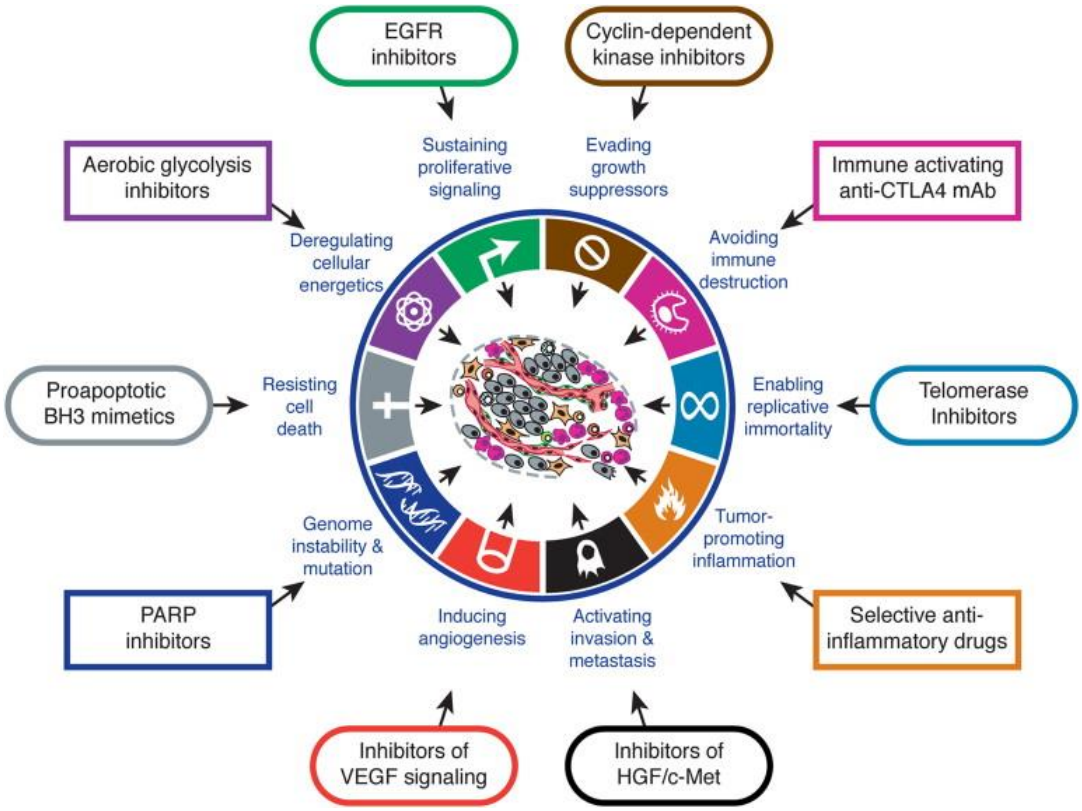
Cancer is a multi-pathway disease, assembled as a heterogeneous and hierarchically organized system (3).

From a molecular perspective, the processes that drive the evolution from a normal cell to a cancer cell involve the sequential acquisition of alterations that damage the cell DNA. The arise of mutations can occur due to endogenous mechanisms such as errors during the replication of the genetic material, due to the chemical instability of the genetic material, or due to the interactions with free radicals produced during the metabolic routes. Likewise, exogenous agents as radiation and chemical carcinogens can be also responsible for damaging the DNA material (18).

A more complex picture emerges when exploring the biological pathways and networks involved in carcinogenesis (19). From a protein level, multiple oncoprotein pathways are implicated in oncogenesis, such as the Receptor Protein-Tyrosine Kinases (PTKs) (20, 21), the Wnt/Wingless Pathway (22, 23), Cadherins cell receptor and Catenin (24, 25), Rho and Ras protein family (26, 27) or the MAP Kinase Pathways (28).

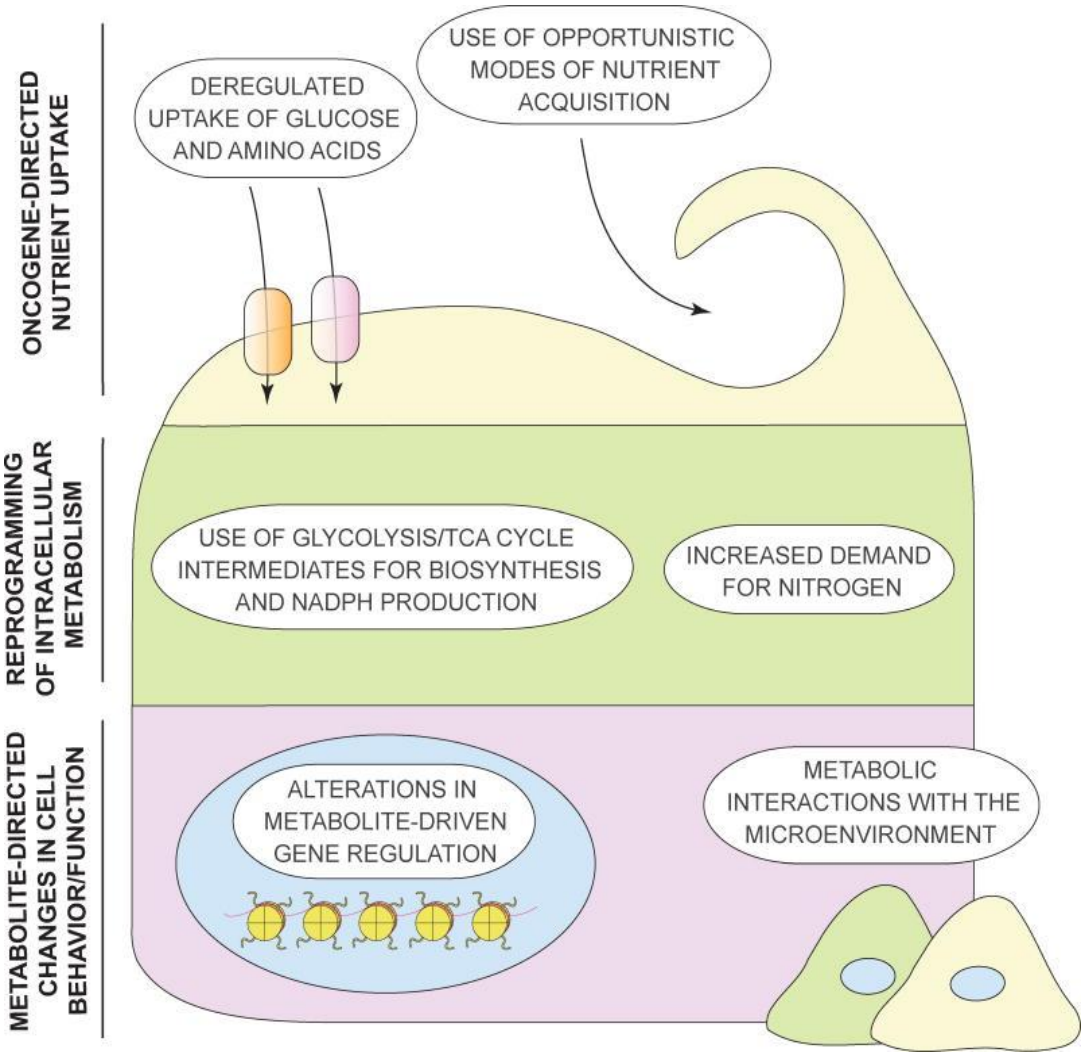
With respect to these pathways, the hallmarks of cancer publication established a new paradigm to understand the development of human tumours (3, 4). Six biological capabilities, upgraded to eight in the more recent publication, were defined as essential when rationalising the complexity of cancer disease. These capabilities comprised, as cited by the authors, the following: proliferating signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism and evading immune destruction. This multistep process establishes the basis of carcinogenesis as displayed in figure 4.

Figure 4. Therapeutic Targeting of the Hallmarks of Cancer. The figure presents the six capabilities defined in the 2000 publication together with the emerging hallmarks and enabling capabilities and the possible therapeutic targets. Figure taken from Hallmarks of cancer. The next generation, 2011 (4).



In 2016, this schema was translated into biochemical pathways by Pavlova and Thompson (29). The hallmarks were deciphered in terms of maintaining the oxygen and nutrients needed to facilitate tumour growth, proliferation and dissemination. The biochemical hallmarks defined were: deregulated uptake of glucose and amino acids, the use of opportunistic modes of nutrient acquisition, the use of glycolysis/TCA cycle intermediates for biosynthesis and NADPH production, the increased demand for nitrogen, the alterations in metabolite-driven gene regulation, and finally, the sophisticated metabolic interactions with the microenvironment, as illustrated in Figure 5.

Figure 5. Hallmarks of Cancer Metabolism. The figure presents the mechanisms utilised by cancer cells to acquire nutrients to ensure cancer cells proliferation and interactions with the microenvironment. Figure taken from The Emerging Hallmarks of Cancer Metabolism, 2016 (29).



This comprehensive framework, comprising molecular, biological and biochemical pathways, illustrates the complexity of the biology of cancer. Defining cancer as a multilayer disease that implies the involvement of diverse biological pathways at different biological levels, including distinct genetic, molecular and clinical events, with pathogenesis varying between patients.

c. Cancer heterogeneity

i. Characterisation of the disease

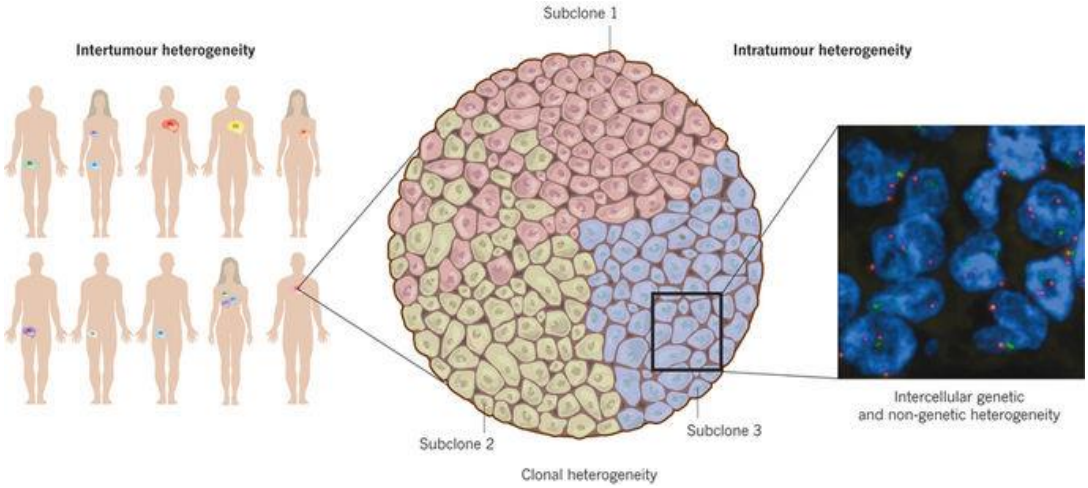
As early as in 1977, the presence of multiple subpopulations in a single mouse mammary tumour was reported. The authors hypothesised that heterogeneity was a general fact in cancer and maybe important for treatment strategies (30).

Heterogeneity is a dissolute term that can describe diverse biological phenomena when applied to cancer. For example, heterogeneity already plays a role in neoplasia, where multiple routes, through the accumulation of genetic and epigenetics alterations, could lead to its differentiation into tumour tissue. However, cancer variation is mainly used to expressed heterogeneity on genetic and phenotypic changes between tumours with different organ and cell origin or between individuals with the same histopathological tumour type, as displayed in Figure 6 (31).

Intertumoral variability, defining heterogeneity between same histopathological tumour types, is highly dependent of patient's characteristics, as age, sex and hormonal or immunological status, therefore population heterogeneity also plays a part in cancer variation. Different phenotypes can occur in hosts with similar characteristics, therefore each patient's tumour may be considered unique while sharing many common features (1). For example, breast cancer (BC) that is highly heterogeneous, may present different morphology, behaviour and clinical implications in patients (32).

Moreover, it is now established that solid tumours can contain different subpopulations of cells with diverse genomic mutations originated from the same primary tumour. This variation within a tumour is identified as intratumoural heterogeneity and is explained as subclonal diversity (33).

Figure 6. Intertumour and intratumour heterogeneity. Figure taken from the causes and consequences of genetic heterogeneity in cancer evolution, 2013 (31).



Given all the previously mentioned elements of disparity, each histopathological tumour type may be considered as a collection of multiple diseases that consist of diverse subtypes. These subclasses that can be defined using different approaches and techniques. The utilisation of immunohistochemistry (ICH) markers, together with clinic-pathological tumour characteristics as morphology, tumour size, volume or grade, has been the classical approach to a clinically relevant cancer molecular characterisation (32, 34). These biomarkers are widely used in standard clinical practice to assess cancer diagnosis, prognosis and treatment selection, but have proven not to be enough to explain all cancer heterogeneity and variation on clinical outcomes (35). Focusing on BC, in the last decades predictive IHC markers including estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) have been extensively used for biological breast tumour subtyping and have been useful to treat highly endocrine responsive tumours but not to treat non-endocrine active tumours that lack of targeted therapies for patients leading to poor prognosis (32, 35).

Therefore, in the last decade with the focus on personalised medicine and the advanced molecular technologies, multiple efforts have been made to reveal a comprehensive molecular characterisation of the disease able to clinically profile the tumours.

In BC, multiple studies have attempted to molecularly classify the disease. In 2000, *C. Perou* published a new classification for the breast cancers, using DNA microarrays and patterns of gene expression. Moreover, in 2012 *Curtis et al*. presented a comprehensive molecular portrait of human BC using 2000 samples (36-38) .

Consequently, heterogeneity should be considered as an inherent characteristic of cancer and unravelling its variation by molecular profiling the disease, could further drive a development in targeted therapeutics interventions and biomarker discovery.

ii. Characterisation of the population: Cancer susceptibility

After exploring the insights of diversity in carcinogenesis, the next step is to understand the heterogeneity that occurs at population level.

Each individual's susceptibility to cancer can be explained as a result of heterogeneity in the following elements: a biological component modulated by genetics, life-style behaviour aspects and environmental factors (39).

Focusing first on **the internal or biological component**, variation in human genetics is vast. The human genome differs in 0.5% between two individuals; a small difference which still implies millions of differences in the DNA (40). The mechanisms responsible for this genetic drift between individuals start with the exchange of genetic material in the sexual reproduction where two haploid cells, formed during meiosis, produce a diploid zygote (41) . At a molecular level, different mutational events arise in genes during a human live. At a population level, natural selection creates positive selective pressure in individuals with a genetic dotation that represents an advantage in the environment adaptation (42).

Most of the genes can present different variants in the population known as genetic polymorphisms. Most common types of polymorphisms are changes in one

nucleotide position such as single nucleotides polymorphism (SNPs), then small insertions or deletions are next type of variation, followed by larger structural alterations (chromosomal rearrangements) and copy number variations in the genome. The Human Genome project, the International HapMap project and the Human Variome project have widely explored genetic variation and identified 10 million of frequent DNA variants, mainly SNPs, in a set of human samples (43-46).

Based on this inherited variation, an individual will have a predisposition to certain cancers due to specific cancer susceptibility gene alterations. Mechanisms of inheritance of cancer susceptibility are the consequence of alterations in different types of genes divided in two main classes, genes that preserve the integrity of the genome called caretakers and genes that control the proliferation cycle or gatekeepers, which completely regulate the tumour growth by suppression of the proliferation or by promoting cell death (47). These two groups of genes include alterations in tumour oncogenes as APC gene, defects on the tumour repair genes as the gatekeepers BRCA1 and BRCA2 mutations associated with susceptibility to breast cancer, colon cancer, ovarian and prostate cancer (47, 48) and others alterations as in the genes that are involved in the stimulation of the vascularisation of tissue (49).

A small proportion of many cancers is due to inherited mutations in above mentioned genes, which results in a high risk of developing specific cancer. Moreover, genes present different penetrance, meaning that the extent of expression of the genes varies in the individuals that carry them. When exploring genetic variations associated with complex diseases, high penetrance genes directly related with disease that confer high risk on the carriers are rare, for example BRCA1 and BRCA2 BC susceptibility genes account for 16-25% of inherited risk of BC. However, lower penetrant genes that are more common polymorphic variants, confer only small risk of disease, which might confirm the hypothesis that the combination of effects of the less penetrant variants together with environmental factors is responsible for the susceptibility to disease (50). This assumption was

confirmed by *Yang et al*. who estimated that the number of genes needed to account for complex disease susceptibility in the population, are between 20 to 50 variants, which would explain only half of the burden of a common disease (51).

Ethnicity also plays an important role in cancer susceptibility. Due to natural selection, populations that are geographically and ancestrally distant tend to differ in their genomic variation expressing diverse polymorphism frequencies (43). This implies that different ethnicities may inherit diverse susceptibility to cancer because of their specific genetic content (52, 53). African-American men have the highest incidence rates for PCa while Japanese men have the lowest (54) . However, the difference of these incidence rates may also be influenced by disparity in cultural, socio-economic, and environmental factors (55-57).

Gender is a well-established factors of cancer susceptibility in the field of cancer epidemiology. Different studies have shown that cancer incidence is higher in men than women, particularly in haematological cancers, sarcoma, lip and larynx (58). Survival rates and mortality rates follow the same pattern for many cancer (59).

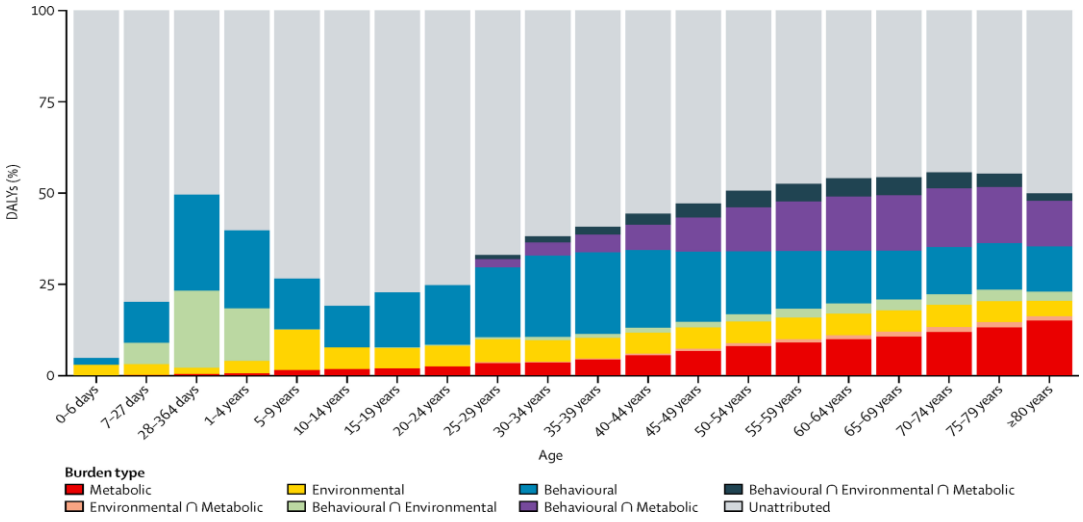
Age has been extensively studied as one of the main risk factors for cancer, given that it is considered as an age-related disease with incidence increasing after midlife. Aging biological processes might be responsible for this association given the accumulation of genetic and epigenetic mutations over time, along with the aggregation of environmental exposures, such as chemicals and radiation and the adoption of health risk-behaviours, such as tobacco use, lack of physical activity, poor nutrition or excessive alcohol consumption (60). Ultimately, age can be considered as a cancer susceptibility risk factor modulated by genetics, social and environmental agents. Moreover, these health risk behaviours are also influenced by one individual's gender (61, 62).

Hence, the following section describes population susceptibility to cancer as a result of differences in **life style behaviours and environmental factors**.

To understand the proportion of attributable risk to behavioural, environmental and metabolic factors, I explored the Global Burden of Disease, Injuries, and Risk Factor study 2013. The study concluded that 50% of global mortality and more than 30% of attributable deaths, years of life lost, years lived with disability, and disability-adjusted life-years (DALYs) can be explained by behavioural, environmental and metabolic factors, with high BMI being one of the leading risk factors (Figure 7).

Moreover, environmental exposures such as air pollution, tobacco smoking and alcohol use are factors that have accounted for about 12.6 million deaths, more than 20% of the global mortality, as reported by the WHO in 2012.

Figure 7. Proportion of global all-cause DALYs attributable to behavioural, environmental, and metabolic risk factors and their overlaps, by age. Figure taken from Global Burden of Cancer, 2013 (16).



Moreover, in 2002 *C. Willett* indicated that non-genetic factors can account for up to 80%-90% of the risk for diseases responsible for high mortality in Western countries, concluding that the most common cancers in developed countries are due to environmental factors (63). He investigated migration studies together with twin studies and estimated that the genetic component of BC was 27%, whereas this was 35% and 42% for colorectal and prostate cancer, respectively (64).

Furthermore, lifestyle factors such as smoking, obesity, unbalanced diet, alcohol consumption and lack of exercise, were confirmed major risk factors for cancer (65, 66).

Smoking increased the risk of lung cancer with a relative risk of 4.4 for men and 2.8 for women, when comparing current smokers with never smokers as presented in a systematic review of the Japanese population (67). Similar results have been observed for cancer of the lower urinary tract in a Germany population (65). Moreover, over 70% of colon cancer has been found to be accountable to smoking habits (63, 68).

Obesity or high BMI have also been associated with various cancers risk such as BC, PCa and colon cancer, probably because an interaction with hormonal pathways (66, 69). In 2012, *Renahan et al.* reported that 3·6% of all new cancer cases in adults were attributable to high BMI, indicating that excess of body weight can be one of the most important preventable causes of cancer, particularly in high-income countries (70-73). High meat intake and high fat diets have also been broadly reported as associated with an increased risk of cancer (68, 74-77). For example, an American prospective study showed elevated risks for multiple cancers (78), on the contrary high intake of fruit and vegetables has been shown to have a protective effect for colorectal cancer (79, 80).

Alcohol intake has recently been confirmed as a definitive risk factor for cancer, accounting for 5%-6% of new cancers and cancer deaths worldwide (81).

Socioeconomic and cultural background are essential when trying to understand the variability of all the above mentioned behavioural lifestyle factors in individuals. For example, lower socioeconomic circumstances are associated with health-risk behaviours and poor cancer outcomes (57, 82, 83). *Danai et al.* presented the leading cancer mortality risk factors for low-middle-income countries versus high income countries. Smoking and alcohol use were significant in both groups, whereas

low fruit and vegetable intake appeared in low-and-middle-income and overweight and obesity were significant in high income countries (84). Ethnicity can be responsible for lifestyle behaviours that may explain differences in cancer risk in different races. For example, it has been shown that the lower BC rates for south Asian and black women in England may be explained by differences in lifestyle and reproductive behaviour (85). Healthy lifestyle behaviours as a balance diet, weight management, regular physical activity and cessation in smoking and alcohol will reduce cancer risk in the population (66).

The **environmental factors**, sometimes measured as occupational factors (e.g. air pollution, heavy metals and various endocrine disrupting chemicals) are also an important component of population cancer susceptibility (63, 86). Rappaport pointed out that given the fact that environmental causes are mainly responsible for chronic diseases such as cancer, the causal agents are still insufficiently characterised (87). The environmental exposures that have been more extensively studied are air pollution, chemicals and radiation.

Environmental chemicals are substances commonly detectable in every day products such as drinking water, foods, dental amalgams and pesticides. These chemicals include heavy metals such as cadmium, mercury and lead, allergens and organic chemicals, for which duration in the environment and body is fluctuating. The mechanisms of action of these toxics in the body varies, with oxidative stress, DNA damage, endocrine and immune disruption, being the main processes associated with cancer disease (88-90). For example, Bladder cancer is a tumour well-known for its environmental component. Risk of bladder cancer is associated with smoking fumes, arsenic in drinking water and occupational exposure to aromatic amines (2-naphthylamine, 4-aminobiphenyl and benzidine) and 4,4'-methylenebis (2-chloroaniline) as a result of chemical dye exposure, rubber and plastics industries (91, 92).

Moreover, exposition to radiation even in low doses induces ionization in the body which causes DNA damage at various levels as damage to single, bases, single-strand breaks, double-strand breaks and multiply-damaged sites. Radiation exposure has been widely associated with BC, established mainly through studies of patients exposed to therapeutic radiation and studies investigating the Japanese bomb survivors (66).

Given all the above-mentioned factors of variability in the population, including the genetic - environmental axis, characterisation of the population can be proven difficult and to fully understand population cancer susceptibility the interplay of all the above-mentioned factors needs to be taken into consideration.

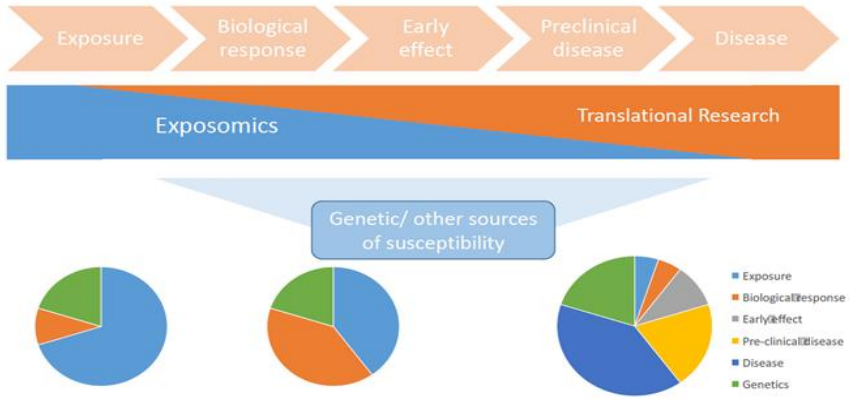
II. Cancer Assessment: Biomarkers & Exposome

Given the above described heterogeneity in cancer susceptibility, the next challenge is to explore different approaches to assess the risk to disease.

Cancer development is a multistep pathway that begins with one individual exposure to an agent that might activate a biological response. This biological process might present early effects in the body that will derivate in the later development of the disease. Therefore, carcinogenesis can be considered a continuum of molecular and genetic alterations leading to disease, as shown in Figure 8. In each stage of the pathway, the susceptibility to disease will depend on different factors in different proportions and consequently, the approach to evaluate the disease will vary along the disease pathway (93-95). For example, in the exposure and biological response stage, exposomics can be an adequate approach to assess disease development, whereas at the disease stage, translational research will be preferable to evaluate disease.

The projects described in this thesis, will consider cancer assessment at the stage of biological response/early effect in the project B and at the disease stage in project C.

Figure 8. Schematic representation of the disease pathway. At different stages of the disease, the susceptibility can be attributable to different factors and the approach to assess the disease will be different. Figure adapted from Molecular Epidemiology of Chronic Diseases and the Exposome Lecture from R. Vermeulen.



a. Biomarkers in cancer

As explored in the previous sections, cancer development involves multiple biological pathways and includes diverse genetic, molecular and clinical events (1, 2). The pathogenesis varies between patients given the broad spectrum of cancer heterogeneity as described above. Therefore, it is critical to identify robust biomarkers that are precisely linked to carcinogenesis to get a better understanding of the complexity of cancer, its causes and evolution, to ultimately improve cancer diagnosis and prognosis (10, 93).

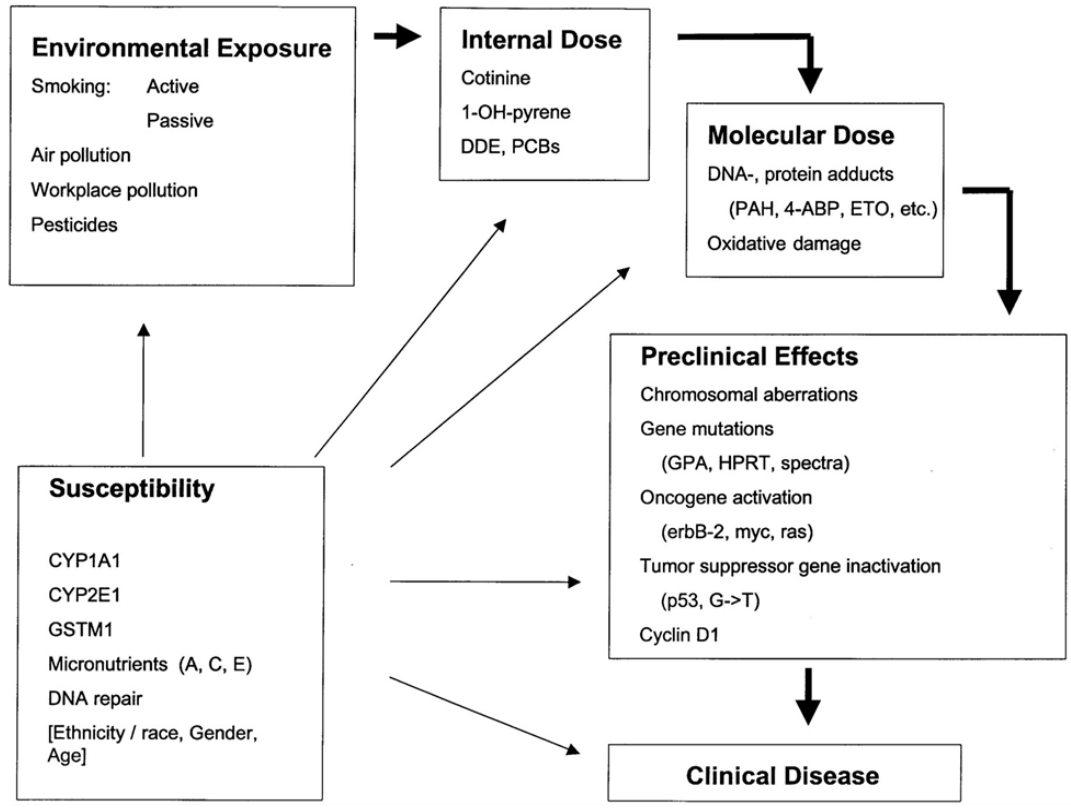
A cancer biomarker or a cancer biological marker, is any measurement taken from a biological sample at any systemic level such as biochemical, molecular, immunological or psychological, that can be monitored and utilised as an accurate indicator of a biological or pathological process or of a therapeutic response in cancer patients (5, 96).

The utility of cancer biomarkers is diverse and encompass the whole spectrum of the disease. Their applicability comprises all disease stages, from the early evaluation of the exposure or causes of disease, the identification of subclinical disease or intermedia endpoints for prevention strategies and the clinically

meaningful classification of disease, to the identification of individuals susceptible to disease for early diagnosis and individuals with high probability of therapeutic response for better prognosis (97).

Perera et al. classified biomarkers following the disease pathway (Figure 9). This classification can be summarised in three main groups: markers of exposure and dose utilise in risk prediction and exposure assessment, markers of effect appropriate for screening, diagnosis and prognosis and markers of susceptibility employ in population stratification and risk assessment (94).

Figure 9. The continuum of biomarkers with representative examples within each category. Figure taken from Molecular epidemiology: recent advances and future directions, 2000 (94).



Nevertheless, biomarkers' clinical relevance will depend on the following characteristics of the marker: precision on the predictive accuracy, measurability,

reproducibility, validity, and in the patients and physician's perception about the use and benefits of the marker (98).

To establish the clinical utility of a biological marker, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are measured. Sensitivity and specificity account for the number of true positive cases and true negative cases detected by the marker. The PPV is the proportion of individuals who tested positive and are true cases and on the contrary, NPV is the proportion of individuals who tested negative and are negative cases. Receiver operating characteristic (ROC) curves are commonly used to the assess the prediction capabilities of a marker by plotting sensitivity values versus 1 - specificity values of the marker in comparison with the values of standard prediction marker commonly stated as "the gold standard" (93, 99). Based on these measurements of performance, a good predictive marker will have high sensitivity and specificity. However, some characteristics might be more desirable depending of the specific used of the marker, for example a diagnostic marker will require a high sensitivity while in a screening marker, both sensitivity and specificity are crucial to avoid the detection of a high number of false positives.

Moreover, to implement a new marker into clinical practice, a careful assessment of the utility of the marker is required. Variability of the measurements compromises the reliability of the marker causing failure of the validation process. Errors in study design and execution of the validation trials, failures on standardisation of the preparation (e.g. specimens' collection, handling and storage of the samples), variation due to intra-interindividual variability and temporal and seasonal variation of the markers are some of the pitfalls in the biomarker validation (93, 97, 100).

Hence, several of the biomarkers that have been studied today in relation with cancer presented poor usability, e.g. CD44, telomerase, mucin 1 (MUC1), mucin 2 (MUC2). However, others are currently used in standard clinical practice, such as

Prostate specific antigen (PSA) in PCa, though its efficacy is being questioned, given the low specificity of the marker and that screening programmes have not shown an effective decrease in mortality rates (98).

During the last decades, advances in imaging technologies together with the new high-throughput 'omics' technology (e.g.: genomics or proteomics) have created many new candidate markers with potential clinical value for cancer diagnosis and prognosis (96, 101) . However, very few of these identified biomarkers have been integrated into clinical practice and their superiority over the "standard" biomarkers such as prostate specific antigen (PSA) or oestrogen receptor (ER), have yet to be proven (96, 98, 102).

This discrepancy can also be seen in data published by the US Food and Drug Administration (FDA) in 2002. The number of publications on cancer biomarkers has rapidly increased, while FDA-approved plasma-protein tests has decreased over the same period of time (96, 103). Furthermore, a report published in Nature in 2011 suggested that high-throughput 'omics' technology has produced more than 150,000 scientific publications on putative biomarkers, leading to only about 100 biomarkers clinically validated (100) .

Thus, despite this increase in data on biomarkers, their translation from bench to clinic has proven to be difficult (100). Moreover, molecular markers often show low values for specificity and sensitivity (5, 8, 9, 96, 104).

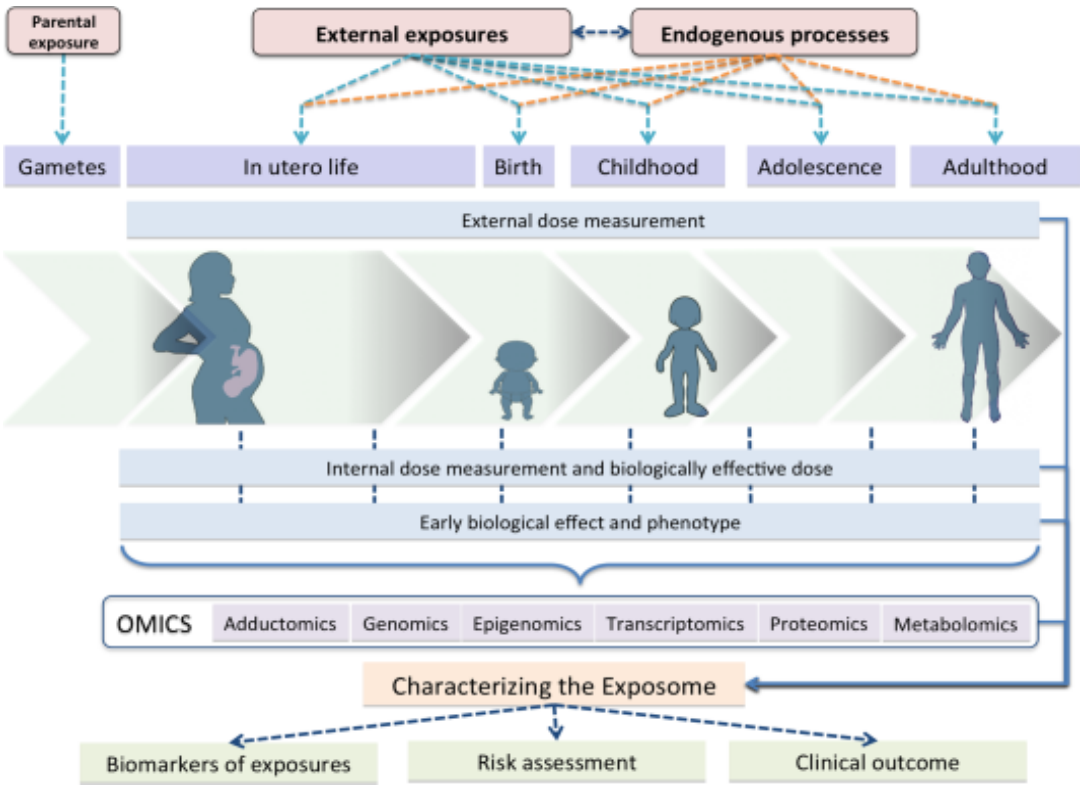
Nevertheless, the challenge of developing cancer signatures through new robust biomarkers is a new exciting research area, given the vast amount of information produced with the rapid advances in omics and imaging technologies. Hence, new approaches to biomarker development have now shifted to systematic-based approaches, replacing single biomarker-based detection by multiple profiling testing simultaneously, which may provide new avenues for biomarker development for cancer disease given the complex nature of the disease (3, 8, 105).

b. Exposome

With the advent of genomics, there has been increasing interest in unravelling the environmental exposure given that the environmental contribution to complex diseases can account for 80 or 90% of the attributable risks (63). Moreover, following the change of paradigm in biomarker discovery for complex diseases, the drift from exploring single markers to more systematic approaches utilising multiple markers concurrently, Professor Christopher Wild introduced the concept of the Exposome in the field of cancer epidemiology in 2005. He defined the concept as ”Exposome is the science of the impact of the environment on health and disease” and refers to the “totality of exposures an individual is subjected to from conception to death” (Figure 10) (106).

Due to the lack of characterisation of the environmental component of disease, further attempts to implement the theoretical concept of the exposome into practice were made. Two main approaches to describe the exposome of an individual during different stages in life were defined: a “bottom-up” concept measuring all the chemicals in each exposure at each time point and a “top-down” idea where all the chemicals or read-outs are measured in a subject’s blood (107).

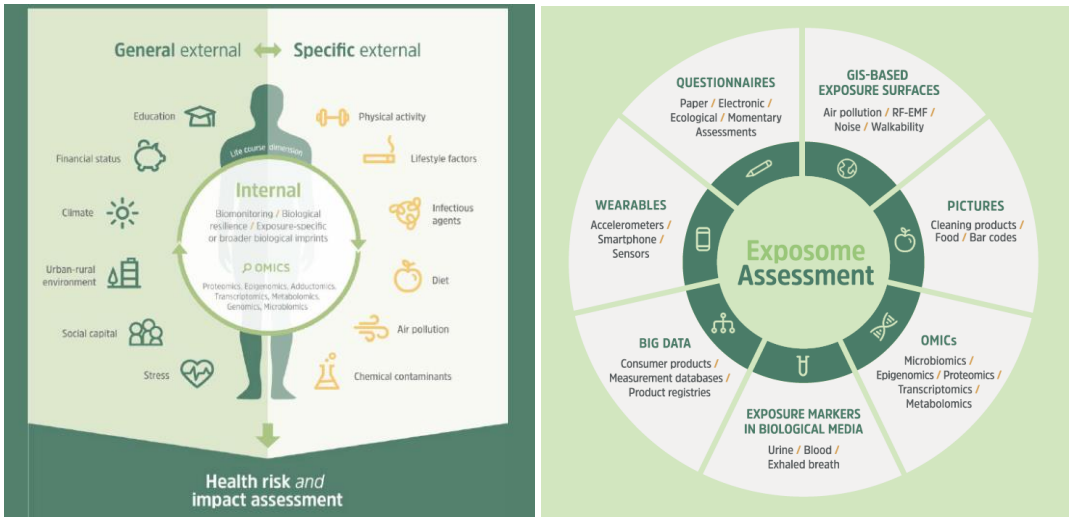
Figure 10. Characterising the exposome. The exposome comprises every exposure to which an individual is subjected over a lifetime. Figure taken from Measuring the Exposome, 2013 (108)



Moreover, the exposome can be represented by three main components that summarise the different exposures that a person may be subject to in different stages of their life: internal, general external and specific external. The internal exposome is represented by the metabolism, endogenous circulating hormones, body morphology, physical activity, gut microbiota, inflammation, and aging; the specific external exposome comprise radiation, infections, chemical contaminants and pollutants, diet, lifestyle factors, occupation and medical interventions and finally the general external exposome accounts for social capital, education, financial status, psychological stress, urban-rural environment and climate (108, 109).

Given the broad spectrum of exposures described above, no one single technology can estimate the entire exposome, a wide range of tools (and statistical techniques) will be necessary to characterise the exposome (Figure 11).

Figure 11. Characterisation of the different component of the Exposome (left) and different technologies available to evaluate each of the components (right). Figure taken from Molecular Epidemiology of Chronic Diseases and the Exposome Lecture from R. Vermeulen.



In 2010, *Patel et al.* conducted a pilot Environmental-Wide Association Study (EWAS) exploring 266 unique environmental factors in relation with Type 2-Diabetes (110). Currently, there are many studies calling on epidemiological research to assess the exposome, with some studies currently collecting data for exposome analysis such as the HELIX project and the EXPOSOMIC consortium (111-117).

i. Blood exposome and metabolites

Rappaport et al proposed a top-down approach to estimate the exposome based on the inherent capacity of the blood samples of characterising the exposome. Blood transports chemicals to and from tissues and represents a reservoir of all endogenous and exogenous chemicals in the body at a given time, and exposures are reflected in the blood given that chemicals are derived from both endogenous

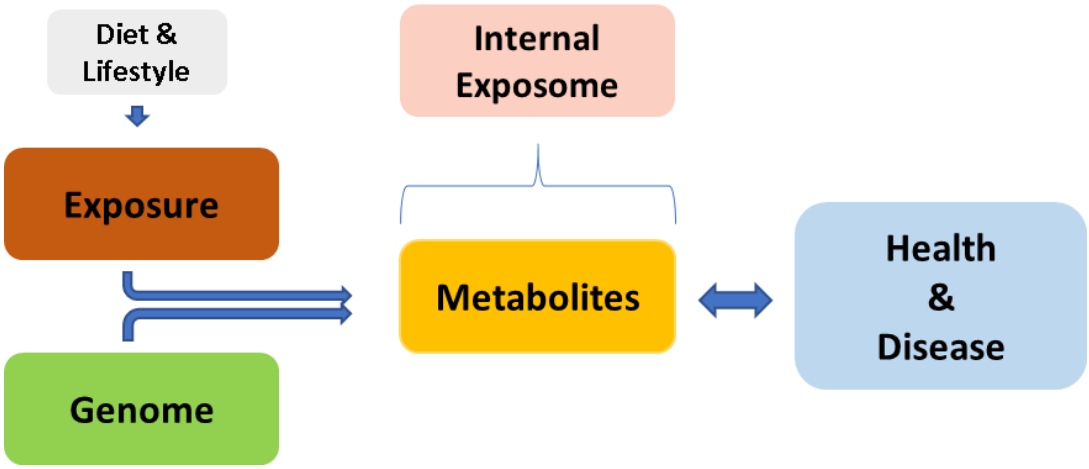
and exogenous origin. *Rappaport* explored 1,561 chemicals (including metabolites, food chemicals, pollutants, drugs) from human samples without finding any order of distinction between endogenous and exogenous chemicals, food and drugs (117).

Therefore, the blood exposome offers an accessible route interrogating all biologically-relevant exposures to identify intermediate or endpoints that can be early biological markers in the causal pathway that links exposure to an etiologic agent for the onset or outcome of disease as suggested by "meet-in-the-middle" approach introduced by *Paolo Vineis and Frederica Perera* in 2007 (95). These markers can be useful to identify subclinical disease and identify susceptible individuals by measuring susceptibility in metabolites for disease stratification (108, 118).

Moreover, metabolites are modulated by genetics and environmental factors, which make them ideal candidates to assess internal and external exposure (108, 119). *Holmes et al.* explored the concept of metabolic phenotypes as a product of interactions among a variety of factors- lifestyle/environmental, gut microbial and genetics, and demonstrated that metabolites discriminate populations for coronary heart disease and stroke based on blood pressure (112).

Figure 12 provides an overview of how the blood exposome encompasses the above described susceptibility factors, establishing blood metabolites as ideal candidates to assess health and disease. The current thesis uses approaches similar to the above-explained concept of the exposome, but specifically focused on a subset of serum biomarkers as part of the internal exposome.

Figure 12. Metabolic profiling and the Exposome. The interplay between the different disease susceptibility components and the metabolites in relation with health outcomes.



Chapter III. Background on statistical methods used to assess disease heterogeneity

This section covers the methods generally employed in cancer studies and specifically focuses on those methods used in the different projects presented in this thesis.

Due to the complexity of cancer disease, the extensive population heterogeneity, and the newly emerged cancer data and demand for high-sensitivity and high-specificity biomarkers, significant sample sizes are required for multiple testing validation. As a result, sophisticated mathematical, statistical and computational multivariate approaches are essential to extract the biological and clinical relevant information related to disease outcome (10, 100).

Therefore, new bioinformatics tools and computation algorithms are being developed to extract meaningful patterns able to predict phenotypic traits and outcome to improve our understanding of cancer biology and ultimately to enhance cancer care (120, 121).

I. Cancer Data

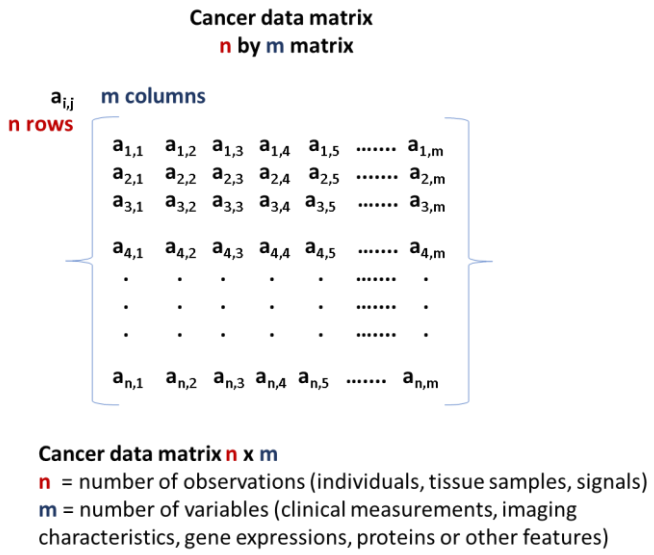
Cancer data is a broad concept that covers information of a wide nature. This data comes in a variety of formats, including numeric and codified values, signals and images, reports, summaries, multivariate time series and concentrations of biological molecules (120).

With the onset of the high throughput technologies following the human genome sequence, Omics data have been produced in vast amounts together with environmental and behavioural population data (45). From the exposome spectrum, cancer external and specific external data contains demographics, environmental, life-style and behavioural factors, while internal data include clinical and OMICS data. Omics data is defined as biochemical measures of the abundance and/or

structural features of molecules involved in important biological processes, such as metabolism and its regulation. The OMICs data follow the central dogma of molecular biology that explains how the biological information flows, where DNA is transcribed to RNA that then is subsequently translated to Proteins (122). OMICS information includes Genomics (DNA material), Epigenomics (DNA) and Transcriptomics (RNA) measuring gene expression and its regulation, Proteomics (Protein) and finally Metabolomics (Metabolites) that characterise the whole cellular activity, including the internal response to external stresses. Ultimately, all these complex interactions among molecular factors, including environmental and behavioural components, will result in the phenotype (121).

Cancer datasets are commonly structured in large data matrices with N observations (e.g. individuals, tissue samples, patients, etc…) and M variables or attributes of each observation (e.g. exposures as clinical biomarkers, imaging characteristics, genes, proteins or other features). Since the utilisation of the high-throughput technologies, M is usually many times bigger than N (Figure 13). When exploring outcome information, a response matrix will be associated with the cancer dataset and a spatial component will be included for longitudinal approaches, increasing the level of complexity of the analysis.

Figure 13. Schematic representation of a cancer dataset.



Cancer data matrix n x m
n = number of observations (individuals, tissue samples, signals)
m = number of variables (clinical measurements, imaging characteristics, gene expressions, proteins or other features)

Cancer data is characterised by different elements: (i) the variability of the dimension of the datasets, going from hundreds to millions (proteins to genes), (ii) the variety of the nature of the variables studied which can be continuous, binary, categorical, counts, etc.., (iii) the noise or error due to experimental conditions and finally, (iv) the heterogeneity of the data given its biological nature. An extra layer of complexity is provided by the underlying correlations that frequently exist between the different features studied. Therefore, the inherent complexity and multidimensionality observed in cancer data requires a flexible statistical framework able to extract meaningful information, whilst identifying the correct associations to predict cancer outcomes (121).

Consequently, different methodologies are required for the exploration and analysis of the data, including data visualisation, information analysis and data mining. Methods like data mining are currently being integrated into clinical practice, while other data reduction methods such as classification and regression models need to be investigated further (120, 123) . Furthermore, these data reduction methods have been broadly used in bioinformatics and computational biology, but there is a lack of integration of biological and clinical data.

II. Methods to reduce data dimension and optimize prediction

Given the dimension of the cancer datasets and the increasing numbers of instances where there are more features that can be predictors of outcome than observations, most of the approaches to analyse this type data aim to reduce the dimension of the datasets, while exploring and extracting the substantial information.

Common avenues for the analysis of the cancer data are different classes of regression-type analysis such as univariate approaches and multivariate approaches that include dimension reduction, feature selection approaches for variable selection and classification/clustering techniques (124-126).

a. Univariate analysis

Univariate analysis explores each of the potential predictors independently in relation with the disease outcome. The relation between the N observations from individuals and the independent predictor in a study can be formulated using a generalised linear model provided that the outcome is a real-valued observation (Figure 14) (127).

Figure 14. Linear model for an observation i and a predictor j. It estimates the association between the predictor of an outcome Y_i and the observation $X_{i,j}$ for a given dataset of sample size=n.

$$Y_i = \alpha + \beta X_{ij} + \varepsilon_{ij} \quad (i= 1,\ldots,n)$$

α = Constant regression coefficient (β_0)
β = Beta regression coefficient or estimated effects
ε_{ij} = Error or noise

Based on generalised linear models, a variety of methods can be utilised to estimate the association between the predictor and the outcome depending on the nature of the outcome.

There are six possible scenarios:

- If the outcome is a continuous variable, we could utilise a linear regression or an ANOVA to estimate an association between the predictor and outcome. For example: association of meat or alcohol intake and glucose levels in blood.
- If the outcome is a categorical variable, we could utilise a multinomial logistic regression, ANOVA or X^2 test to estimate an association between the predictor and outcome. For example: association of meat or alcohol intake and low, normal and high glucose levels in blood.

- If the outcome is a binary variable, we could utilise a logistic regression, T-test or X^2 test to estimate an association between the predictor and outcome. For example, presence of glucose levels in blood and death of bladder cancer.
- If the outcome is a count variable, we could utilise a Poisson regression to estimate an association between the predictor and outcome. For example, glucose measurements in blood and risk of bladder cancer over a period of time.
- If the outcome is a time, we could utilise a survival model to estimate an association between the predictors and outcome. For example: glucose levels in blood and time to bladder cancer-specific death.

The different methodologies mentioned above, estimate the relation between the predictor and outcome following different approaches:

- **Linear, logistic, multinomial, Poisson regression and survival** analyses model the relationship between respectively, a scalar, binary, categorical or time outcome variables and one or more explanatory variables (or independent variables) (127, 128).
- **T test and chi-squared test X^2** test explore the sampling distribution and determine if two or more sets of continuous data or frequencies are significantly different from each other (129, 130).
- **ANOVA** analyses the variance or differences among different group means (129, 131, 132).

This univariate approach estimates each feature of the observation as an independent predictor of the outcome studied, which implies multiple testing to identify true associations with disease outcome. The significance of the association between the predictor and outcome is measured by comparing the result of the chosen test to the $(1-\alpha)^{th}$ percentile of the distribution of the statistic under the null hypothesis that implies no association. Type I error α defines the chance of rejecting the null hypothesis when it is true (false positives). The type I error or false

positive error will increase with the number of tests taken to test the hypothesis in the dataset; hence, correcting for multiple testing needs to be considered when performing multiple univariate analyses in a dataset.

There are several approaches to assess **multiple testing**, for instance by controlling for the Family-Wise Error Rate (FWER) or by regulating the False Discovery Rate (FDR). Based on these two routes, there are multiple methods available such as *Benjamini-Hochberg, Bonferroni, Holm's* and *Šidák* corrections that can be employed for multiple testing correction (133).

Mixed models, allow for a greater flexibility to model the association between exposure and outcome given that both linear and non-linear models can be mixed together. This can be advantageous when prior biological knowledge about the association exists, allowing to account for random effects due to the variance of observations (134, 135).

Another standard univariate approach commonly used is **correlation analysis**. To understand the possible interactions between the two continuous features in the cancer dataset, correlation coefficients are usually performed (136). There are two main type of tests available to measure the possible two-way association between two continuous features which use depends on the underlying distribution:

- *Pearson's correlation coefficient* which quantifies to which extent there is a linear relationship between the variables.
- *Spearman's correlation coefficient* which quantifies to which extent there is a monotonic relationship between the variables (not necessarily of a linear nature) which one variable can rise or drop while the other increase.

To interpret the direction and intensity of the interaction, the *r coefficient* is used. *r coefficient* is dimensionless and it can range from +1 to -1. The r coefficient can be interpreted as indicate below:

- r = ± 1 indicates a perfect linear (Pearson) or monotonic (Spearman) association.
- r = 0 indicates no relationship.
- r < 0 indicates a negative relationship. The more negative the coefficient, the stronger the association.
- r > 0 indicates a positive relationship. The more positive the coefficient, the stronger the association.

As explored previously, univariate approaches are computationally efficient and adaptable to different outcomes, although the potential of simultaneous predictive effect between different predictors and the outcome is ignored, lacking of a systemic approach to the disease studied (137).

b. Multivariate Analysis

Multiple factors interplay in biological disease mechanism, clinical outcomes or population risk stratification, as explored in previous sections. Therefore, multivariate analysis appears to be the adequate approach to study cancer risk. Multivariate approaches will explore multiple potential predictors and their correlations in relation to disease simultaneously, given that different predictors may explain diverse interactions with the outcome of study.

The two main avenues for multivariate analysis, while reducing the dimension of the cancer multidimensional data matrix, consists of either grouping the N observations or the M features into a number of G clusters or selecting a number of features that contain the main variance of the data in relation with the outcome of study. Dimension reduction techniques, feature selection approaches for variable selection, and classification/clustering techniques are the standard methodologies performed for this type of analysis. Data reduction methods can thus be defined as mathematical algorithms that decrease the data dimensionality to help its exploration, by summarising the information into a smaller number of components that represent the main structure of the data (5, 9).

Dimension reduction techniques

These methods reduce the number of features or variables in the multidimensional matrix into a set of variables of lower dimensionality by applying (usually linear) transformation to the data. The new variables or components are linear combinations of the original features that produce a reduced representation of the original data. The linear combinations are computed to capture the variability of the original data by optimising a specific measure of the mathematical distance between the original and transformed feature values (138). These methods can be categorised as feature extraction methods.

Moreover, these methods can be unsupervised (where no prior information of the data structure is known) or supervised (where information of the structure and correlation with the outcome of interest is considered). **Principal component analysis** (PCA) is a well-established unsupervised method that reduces data by identifying directions based on the dominant eigenvectors of the data covariance matrix, called principal components, projecting the data onto the dominant eigenvectors ensures that the variation of the original features is maximal (139, 140). It has been broadly employed in clinical diagnosis, such as radiopharmaceutical assays to evaluate organ function (141) and to weigh up the expression of genes in tumour samples (142).

Another similar technique is **Factor analysis** (FA) which also models the observed variables as linear combinations of potential factors that will define the different classes available in the data. However, this technique considers random errors in the measurements, optimises the common variance of the data and assumes that the observed variables were the linear combinations of a smaller number of underlying and unique factors (143-145). Factor analysis can either be supervised (termed confirmatory analysis) or unsupervised (termed exploratory analysis). In confirmatory factor analysis, a hypothesised factor structure is tested using the data whereas exploratory factor analysis does not require any prior information about the classification. For instance, exploratory factor analysis has been employed to

cluster biomarkers to explore their associations with cardiovascular disease (146) and confirmatory factor analysis has been utilised to confirm the hypothesis that the components of the metabolic syndrome are manifestations of a single common factor (147, 148).

Another technique widely used in biological studies is the **Partial Least Square regression** approach (PLS) (149). This supervised method explores latent variables that are capable of constructing the outcome variables, so that correlation with the outcome is explored in this method. This regression approach implies an a priori assumption of the number of latent variables that will define the data. This method has been used in cancer research, for example to discriminate between benign and cancer samples based on gene expression profiles (150).

The above-mentioned dimension reduction methods have proven to be useful to reduce the dimension of the data, aiding in the exploration and analysis of the data, however sometimes the latent components are difficult to interpret from a biological perspective.

Feature Selection methods

To overcome the limitations of the above methods, alternative variable selection techniques have been explored to reduce the number of features by selecting a combination of predictors in a cancer dataset that best define the outcome. These approaches are developed mostly for linear and nonlinear regression models in which the impact of the predictors is mediated strictly via a linear combination as in the model of Figure 14. They all involve modifying maximum likelihood estimators of the regression coefficients by adding penalty terms (that favour null regression coefficients).

The two-main variable-selection techniques via penalty terms are **LASSO regression** and **the Elastic net**. All these methods are regressions applied to analyse multiple regression data that contains multicollinearity i.e. data that exhibit linear mutual relationships.

To account for more complex variability in the associations between exposures and outcome, more general Bayesian approaches of these methods have been developed.

c. Data reduction methods applied in this Thesis

This section provides more background about the data reduction methods employed in the projects of this thesis which belong to the clustering/classification methodologies.

Latent-class analysis, a specific type of cluster analysis

Latent-class cluster analysis (LCA) is a specific type of model-based clustering and is the method selected for one of the projects performed in this thesis. Clustering is a data reduction tool that has been widely used in the context of big data. Related samples are clustered according to similarity or dissimilarity coefficients. Hence, clustering reduces data to a manageable size allowing easier identification of associations and patterns. Similar to variable selection approaches, these methods are considered feature selection methods.

Classification or clustering methods also play an important role in medicine. Patients can be classified based on some variables, which can be symptoms, clinical or physical measurements, or response to treatments such as surgery. Classification of patients can be useful for statistical diagnosis, medical prognosis and prediction of cancer outcomes (151). As a result, the classification problem in medicine has been approached from different angles, resulting in a wide variety of methods available and are commonly used in bioinformatics. It can be applied to cluster genes or samples with similar expression patterns. Many clustering algorithms exists such as K-means, density-based algorithms, general type-2 fuzzy sets and probabilistic clustering (152, 153).

Overall, clustering is a classification of observations into groups on the basis of some specified similarity criterion. **Unsupervised clustering** methods carry out the

classification without making use of a priori knowledge of any group structure (even when such information is available). Here, determining the optimal number of clusters required to fit the data is a difficult task. In contrast, in **supervised clustering** one knows and uses a priori information on the class membership of each sample, and the task is to extract from this information a protocol for assigning as yet unclassified samples reliably to their classes. This is also called **discriminant analysis**. Each observation belongs to one of the specific groups, of which the number is by definition known. This a priori knowledge of the number of groups reduces the complexity of the cluster analysis.

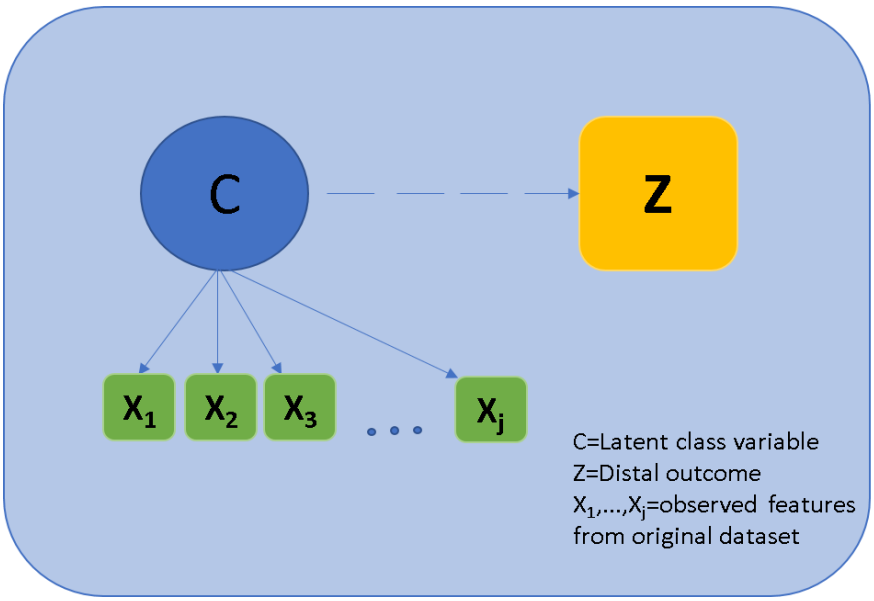
There are many types of clustering algorithms available, each with their own specific features, based on the different assumptions made to evaluate the groups that represent the data and on the (implicit or explicit) measures used to quantify similarity. For example, the classical methods, hierarchical and partitioning clustering, are heuristic methodologies, where classification is based on the nature of the data (e.g. similarity, dissimilarity between variables). This implies that the class structure is not known beforehand, which makes the choice for a final cluster approach ad-hoc and subjective. In contrast, **model-based clustering algorithms** are probabilistic by nature, based on an assumed generative model for the population, offering an alternative to the heuristic algorithms. In the model-based approach, the problem of determining the number of clusters is replaced by the challenge of choosing the right generative model for the data. This method has been tested in comparison with the established heuristics algorithms in gene expression studies, showing a superior performance, consistently selecting the correct model and number of clusters (154).

Latent Class Cluster Analysis (LCA) is a specific type of model-based clustering and is the method selected for project B in this thesis. This clustering method is commonly used in social sciences (13, 155). The goal of this analysis is to use a representative community sample rather than one individual sample to characterise different classes, thus facilitating interpretation of data. However, very few studies

have used LCA to explore underlying relationships between biomarkers and disease outcomes. This method differs from the clustering algorithms listed above in that instead of grouping biomarkers, it groups subjects according to the latent class under which they are categorised (156).

LCA is a statistical method for the analysis of multivariate data. It was developed to handle only categorical data types (156), but later on improvements were made to allow use of continuous data and mixed data types (157). The aim of this algorithm is to stratify the observed (manifested) variables by an unobserved "latent" variable that eliminates the confounding between those variables which are assumed to be mutually independent. The model will group each observation from the dataset probabilistically into a "latent class", which could explain how that observation will respond to the manifested variables. The relationship between the manifested variables can thus be explained in terms of a latent variable (156) . This idea is illustrated in Figure 15, where X_1: X_j manifested variables are explained by the latent variable C that can be associated with an outcome of interest Z.

Figure 15. Schematic representation of Latent Class Cluster Analysis. Manifested variables of a dataset can be explained in terms of a latent variable, associated with an outcome of interest Z.



In mathematical terms, LCA is a statistical model for a sampled population. This model assumes that a heterogeneous population comprises multiple homogeneous subpopulations (named as classes, groups or clusters) with a different multivariate probability distribution function. Therefore, the final model will be a finite mixture of distributions (158). More specifically, LCA is a form of model-based clustering, which assumes that the data is generated by a finite mixture of underlying probability distributions such as multivariate normal distributions or Gaussian mixture (159). Therefore, the clustering problem becomes a matter of estimating the model parameters of the different distributions.

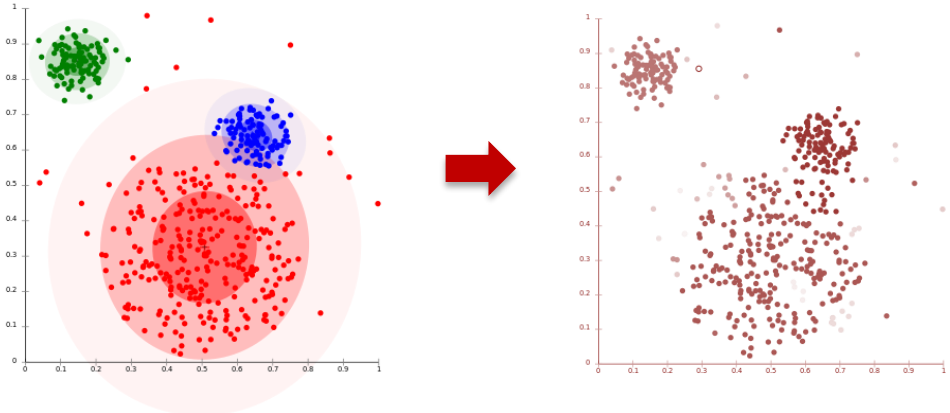
Firstly, we can define the basic latent class cluster model by the following formula:

Figure 16. Formal mathematical formulation of the Latent Class Analysis.

$$P(y_n \mid \theta_{1,..,S}) = \sum_{j=1}^{S} \pi_i P_i (y_n \mid \theta_i)$$

where Y_n is the n^{th} observation of the manifest variables, S is the number of classes, and πj is the prior probability of membership in class j. $P_j (Y_n \mid \theta j)$ is the class-specific probability of Y_n given the class specific parameters θj. P_j is a probability mass function when the manifest variables are discrete and a density function when the manifest variables are continuous. For continuous data, the cluster specific parameters to be estimated are therefore the mean and variances (156, 160).

Figure 17. Schematic representation of the fitting of the data into a mixture of Gaussian distributions followed by the LCA method. The LCA analysis fits the data into a finite mixture of underlying probability distributions such as multivariate Gaussian distributions represented in different colours in the below figure.



Thus, when using the LCA model one needs to estimate of all the above probabilistic parameters of the class-dependent Gaussian mixture models, which is computationally nontrivial (Figure 17). This is usually done via the Maximum A Posteriori Probability (MAP) protocol, using the Expectation-Maximization (EM) algorithm and Newton-Raphson (RM) based search algorithms to find the most probable point. The algorithms iterate until convergence to a maximum global optimum, whereby the MAP estimates of the model parameters (mean, covariance and latent class mixing proportions) and a class-specific prior probability can be obtained. Once these parameters are estimated, one has to decide on the best possible LCA model and the optimal number of latent classes. Ideally, one chooses a model that is complex enough to fit of the data, but avoids overfitting. The *Bayesian Information Criterion* (BCI), the *Akaike's Information Criterion* and the *Chi-squared (X^2)* are broadly used statistical tools for evaluating a set of competing LCA models; these criteria penalise on the number of model parameters to avoid overfitting of the data (158, 159).

Overall, the LCA model has multiple benefits in comparison to other clustering methods. It has the flexibility and possibility to handle very complicated distributional forms, provides a less arbitrary estimation of the number of clusters, and there is a formal criterion to make a decision about the number of clusters and parameters to include. It also avoids the unnecessary scaling of the observed values when the distribution is not normal and it has the possibility of using different data types. Finally, LCA is a probabilistic approach, such that it also takes into account the uncertainty about the object class membership, which can be very useful when interpreting the results (161).

Given the complexity of LCA as a model-based clustering method, several statistical software packages have been developed to estimate its parameters. For the projects B, I have used statistical packages developed for SAS and R. *PROC LCA* was developed for SAS to conduct LCA, multiple-group LCA, and LCA with covariates

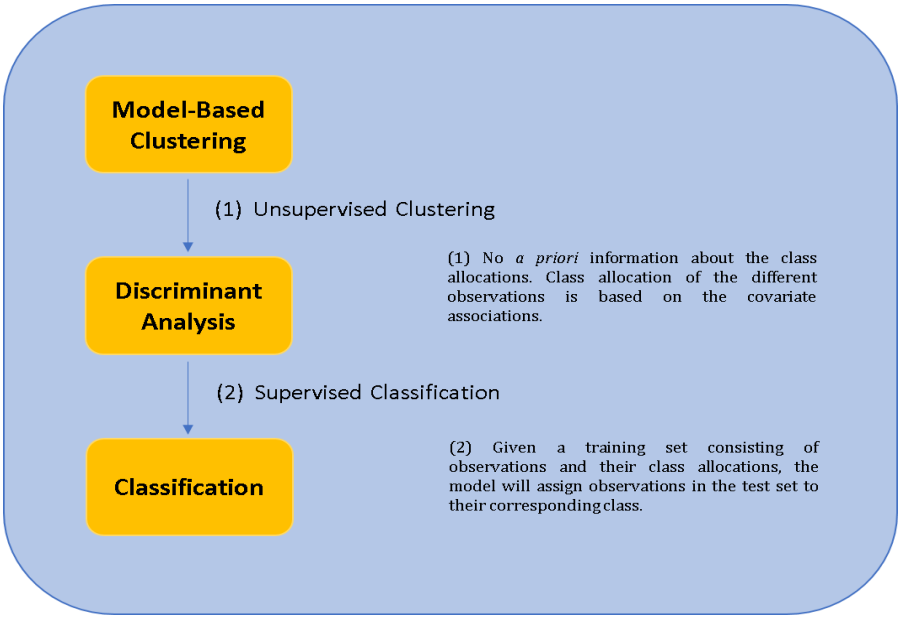
using categorical data (162). *poLCA* and *MCLUST* are R software packages (163, 164). *poLCA* performs a latent class analysis using categorical variables, whereas *MCLUST* estimates density functions that permit the use of continuous variables (157, 165). *MCLUST* automatically estimates the best mixture model according to different covariance structures and different numbers of clusters. It can carry out unsupervised and supervised clustering (Figure 18).

Naïve Bayes Classifier

Another method based on model-based probabilistic approach is the Naïve Bayes Classifier, which is used in project C of this Thesis. The Naïve Bayes Classifier performs discriminant analysis in high-dimensional data that can be useful to predict clinical outcome for future patients (Figure 18). This method is an optimisation of the LCA method and it was developed to overcome the heavy computational demand of the Bayesian probabilistic approaches whilst avoiding overfitting.

This Bayesian latent class regression enables reliable extraction of individualized predictive patterns from cancer data with much larger dimensions and increased accuracy, compared to what is feasible with standard regression and machine learning tools, enabling the identification of statistically significant disease or host heterogeneity in multivariate associations (166).

Figure 18. Schematic representation of the MCLUST and Naïve Bayes Classifier method pipeline.


Model-Based Clustering
(1) Unsupervised Clustering
Discriminant Analysis
(1) No *a priori* information about the class allocations. Class allocation of the different observations is based on the covariate associations.
(2) Supervised Classification
Classification
(2) Given a training set consisting of observations and their class allocations, the model will assign observations in the test set to their corresponding class.

Chapter IV. Population susceptibility to Disease.

To meet the cancer challenge, understanding the insights of disease and population heterogeneity will be fundamental pieces on the cancer puzzle, specifically towards the pursuit of personalised medicine. Comprehending all sources of variability will allow for an accurate and effective stratified medicine.

As highlighted in the previous chapters, cancer datasets have grown extensively in the last decades, due to high throughput technologies and to the need of big samples sizes to establish new biomarkers. Moreover, single biomarkers or single clinical symptoms seem insufficient to capture patients' heterogeneity and tailor treatments, therefore more systematic efforts exploring multiple markers with advance statistical methods are necessary to improve early diagnostic tools in cancer disease. However, to build efficient stratification models for diagnosis, the unnecessary information that increases complexity and noise should be discarded by reducing the dimension in the phenotype space to its predictive components.

Thus, this chapter aims to explore multiple standard of care serum markers in relation to cancer diagnosis following a two-step approach: Firstly, understanding the role of these markers as a subset of the blood exposome and secondly, characterising the potential of these standard of care established markers as markers of cancer susceptibility. The molecules studied in this chapter, blood metabolites, both modulated by environment and genetics, are considered adequate markers to investigate a subset of the exposome and its association with cancer disease and mortality.

Hence, in this chapter I investigated population susceptibility to cancer in a clinical setting, exploring whether data-driven approaches could develop effective cancer risk stratification tools using the Swedish Apolipoprotein MOrtality RISk study (AMORIS) cohort in two different projects:

a. Project A: Blood exposome in AMORIS:

   I considered the study of a subset of the blood exposome by (1) evaluating interactions, as correlations, in routinely assessed health biomarkers and (2) assessing how the external environment influences these interactions.

b. Project B: Metabolic profiles as risk factors in AMORIS:

   I evaluated how multiple markers of blood exposome, when reduced to metabolic profiles, can assess risk of cancer and mortality in a given population.

The projects covered in this chapter are based on data available from the Swedish AMORIS database, so that the first section of this chapter describes this valuable data resource in more detail. The second section of this chapter outlines the health serum markers studied as exposures in both projects. Next, each specific project is explained in more detail.

I. Data Source: AMORIS

The Swedish Apolipoprotein MOrtality RISk study (AMORIS) is one of the largest prospective cohort studies worldwide with detailed information on serum biomarkers. Between 1985 and 1996, the Central Automation Laboratory (CALAB) processed fresh blood and urine samples from 812,073 individuals on a diverse number of markers, comprising 49% men and 51% women. CALAB, a leading centre for collection and analyses of blood and urine samples in the Stockholm county, processed analysis from individuals who were either healthy individuals referred for clinical laboratory testing as part of a general health check-up through occupational health care or outpatients. The results from those analyses were donated in 2002 to the Karolinska Institutet, Stockholm, Sweden, for research purposes. The AMORIS cohort was initially set up to study apolipoproteins and the risk of fatal stroke and infarction. This database with information on more than 500 biomarkers, has been linked to several Swedish national registries including the National Cancer Register,

the Patient Register, the Cause of Death Register, the consecutive Swedish Censuses during 1970-1990, and the National Register of Total Population by using the Swedish 10-digit personal identity number (Figure 19A, 19B). These linkages provide detailed information on demographics, lifestyle, socio-economic status, vital status, cancer diagnosis, comorbidities and emigration. Data from the National Cancer Register provides information about cancer diagnosis prior to assessment of blood biomarkers in CALAB (index date) as well as incident cancer after this index date. AMORIS has been updated with extended follow-up until 31st December 2012 and enriched with clinical cancer information from the detailed Swedish Breast, Prostate, and Colorectal Cancer Clinical Quality Registers. Up to 31st December 2012, 153,820 deaths (18.9%) and 144,533 incident cases of cancer (17.8%) were identified. The AMORIS study conformed to the declaration of Helsinki and was approved by the ethics board of the Karolinska Institute (167).

The following table presents the characteristics of the AMORIS population in comparison with the general population of the Stockholm county (Table 1) (168).

Table 1. Sociodemographic characteristics of subjects in the AMORIS cohort at time of inclusion (1985-1999) compared to the general population of Stockholm County in 1990.

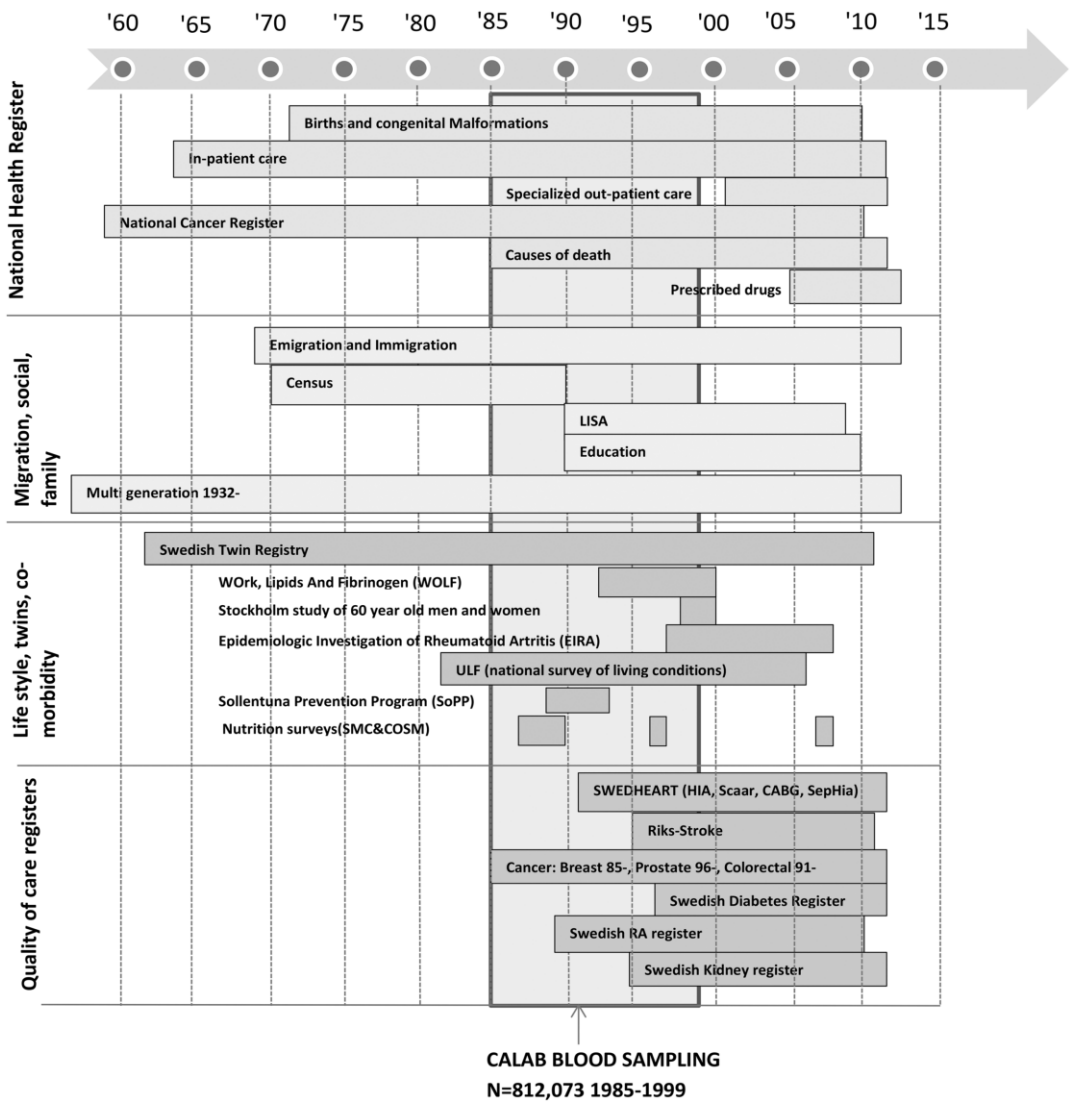
	AMORIS at baseline	**Stockholm county (1990)**
Number of subjects	812,073	1,654,766
Female	51%	52%
Mean age (years)	42.6	38.4
Age distribution		
≤20 years	8%	24%
20-39 years	39%	31%
40-59 years	38%	26%
60-79 years	13%	16%
≥80 years	2%	3%
Country of birth		
Sweden	85%	84%
Finland	5%	5%
RoW	10%	11%
=< 9 years education	28.3%	24.3%
Married	43.3%	36.5%
Gainfully employed (20-64 years)	89.8%	85.1%
Socioeconomic group		
Unskilled worker	90,680 (20%)	177,362 (22%)
Skilled worker	58,598 (13%)	117,969 (15%)
Lower employee	108,315 (22%)	165,185 (20%)
Intermediate employee	110,329 (23%)	174,235 (22%)
Higher employee	89,918 (18%)	135,042 (17%)
Self employed	15,446 (3%)	35,691 (4%)
Farmer	3,377 (1%)	1,437 (0%)

Figure 19A & 19B. Different databases linked to AMORIS. Figure taken from Cohort Profile: The AMORIS cohort, 2017 (167).

A)

Database	Coverage	N	Outcome	Demography	Lifestyle	Comorbidity
Clinical Cancer Quality Register	1997-2012	Breast: 14,934 – Prostate: 17,141 – Colorectal: 6,358	•			
Swedish Cancer Register	1958-2011	148,364	•			
Cause of Death Register	1961-2011	153,800	•			•
National Patient Register	1964/87-2011	Outpatient: 683,747 – Inpatient: 696,825	•			•
Total Population Register	1968-2012	800,587	•	•		
Census Data	1970-1990	783,922		•		
LISA database	1992-2010	799,647		•		
Medical Birth Register	1973-2011	204,449			•	
Multigeneration Register	1932-2011	812,073			•	
National Diabetes Register	1996-2011	58,985			•	•
Prescribed Drug Register	2005-2012	681,299			•	•
Karolinska Institutet (5 Research Cohorts)	1963-1990	29,000		•	•	•
AMORIS cohort (Biomarker measurements)	1985-1996	812, 073				

B)



Currently, 104 scientific manuscript have been published based on the AMORIS cohort, mainly in the area of cancer and cardiovascular disease. Some of the publications investigating the association between serum biomarkers and cancer are described in Table 2 (169-184).

Table 2. Different studies exploring association between serum markers and cancer in the Swedish cohort AMORIS.

Author	Cohort	Biomarkers	Outcome	Results
Van Hemelrijck, M et al.(169)	24,820	Immunoglobulin E	All cancer types	No clear association.
Van Hemelrijck, M et al.(170)	102,749	C-reactive protein & Leukocytes	All cancer types	Positive association between inflammatory markers and cancer risk using repeated measurements.
Van Hemelrijck, M et al.(171)	200,660	Cholesterol, Triglycerides & Glucose	Prostate Cancer	Negative association between glucose and prostate cancer risk. Positive association between hypertriglyceridemia and prostate cancer risk, in combination with high glucose levels. No association for hypercholesterolemia.
Van Hemelrijck, M et al.(172)	545,460	Gamma-glutamyl Transferase, Alanine amino Transferase, Glucose & Triglycerides	All cancer types	A positive association was found between GGT and overall cancer risk. Glucose increased the association for certain cancers. No effects of ALT or triglyceride levels on risk were found.
Van Hemelrijck, M et al.(173)	69,735	Apolipoprotein A-I, Apolipoprotein B, Triglycerides, Total Cholesterol, Glucose, LDL & HDL	Prostate Cancer	ApoA-I and HDL were inversely associated with PCa risk.
Melvin, J et al.(174)	234,494	Lipid profile & Glucose	Breast and Ovarian Cancer	A weak protective association was found between levels of triglycerides and risk of breast cancer.

Seth, D et al.(175)	225,432	Glucose, Triglycerides & Total Cholesterol	Endometrial Cancer	Total cholesterol, triglycerides and TG/HDL ratio were positively associated with EC risk.
Van Hemelrijck, M et al.(176)	542,924	Lipid profile, Glucose & BMI	Kidney Cancer	Triglycerides was positively associated with kidney cancer risk.
Van Hemelrijck, M et al.(177)	196,022	Calcium & Albumin	Prostate cancer	The weak negative association between Calcium and PCa risk is likely to be explained by the relation between Calcium and death.
Wulaningsih, W et al.(178)	540,309	Lipid profile	Gastrointestinal cancers	TC is associated with rectal cancer risk as well as TG with oesophageal and colon cancer risk.
Gaur, A et al.(179)	220,642	Serum Iron, Total-Iron Binding Capacity & C-Reactive Protein	All cancer types	An inverse relation between TIBC and cancer risk was found.
Wulaningsih, W et al.(180)	397,292	Inorganic phosphate	All cancer types	A higher overall cancer risk with increasing Pi levels in men and a negative association in women was found.
Wulaningsih, W et al.(181)	492,044	Calcium & Albumin	Gastrointestinal cancer	A positive relation between serum calcium and oesophageal and colorectal cancer was found.
Wulaningsih, W et al.(182)	11,998	Serum Glucose & Fructosamine	All cancer types	A positive trend was observed between standardized log overall mean glucose and overall cancer risk. Including standardized log fructosamine resulted in a stronger association and an inverse association between fructosamine and cancer.
Arthur, R et al.(183)	14,294	Serum Glucose, Triglycerides & Total Cholesterol	Prostate Cancer	High serum levels of glucose and triglycerides were associated with PCa aggressiveness and severity.
Ghoshal, A et al.(184)	205,717	C-Reactive Protein, Albumin & Haptoglobin	Testicular and penile cancer	No association found between the markers and Testicular and penile cancer.

*The following abbreviations have been used in the table: Total Cholesterol (TC), High-density lipoproteins (HDL), Low-density lipoproteins (LDL), Triglycerides (TG), Apoliprotein A-1 (Apo A-1), Alanine Amino Transferase (ALT), Gamma-Glutamyl Transferase (GGT), Total Iron Binding Capacity (TIBC), Body Mass Index (BMI), Phosphate (Pi), Prostate Cancer (PCa) and Endometrial Cancer (EC).

II. Exposures: Blood metabolites from AMORIS

The laboratory analyses included in the AMORIS study were performed on fresh blood and urine samples by CALAB. The CALAB sample set includes 35,815,102 laboratory values recorded for 595 biomarkers, including mainly chemistry and haematology viral serology biomarkers, but also immunology, allergy and bacteriology markers are available in smaller numbers. Most of the markers measured are standard of care markers, however there are some novel markers such as Apolipoproteins (A-1 and B) and Fructosamine (168). All serum biomarkers were analysed using fully automated multichannel analyser Technicon DAXTM 96 Multichannel Analyzer (Bayer Diagnostics, Tarrytown, USA), using the methodologies described in Table 3 for some of the markers studied in this thesis. More detailed on the laboratory techniques performed in CALAB can be found elsewhere (185-188).

The serum markers studied to investigate population susceptibility in this chapter, were selected to represent the main metabolic pathways, at cellular and organ level. These markers are not necessarily related to cancer outcome, as illustrated in Figure 20, but allow for an exploratory approach based on a set of routinely collected serum biomarkers available in AMORIS.

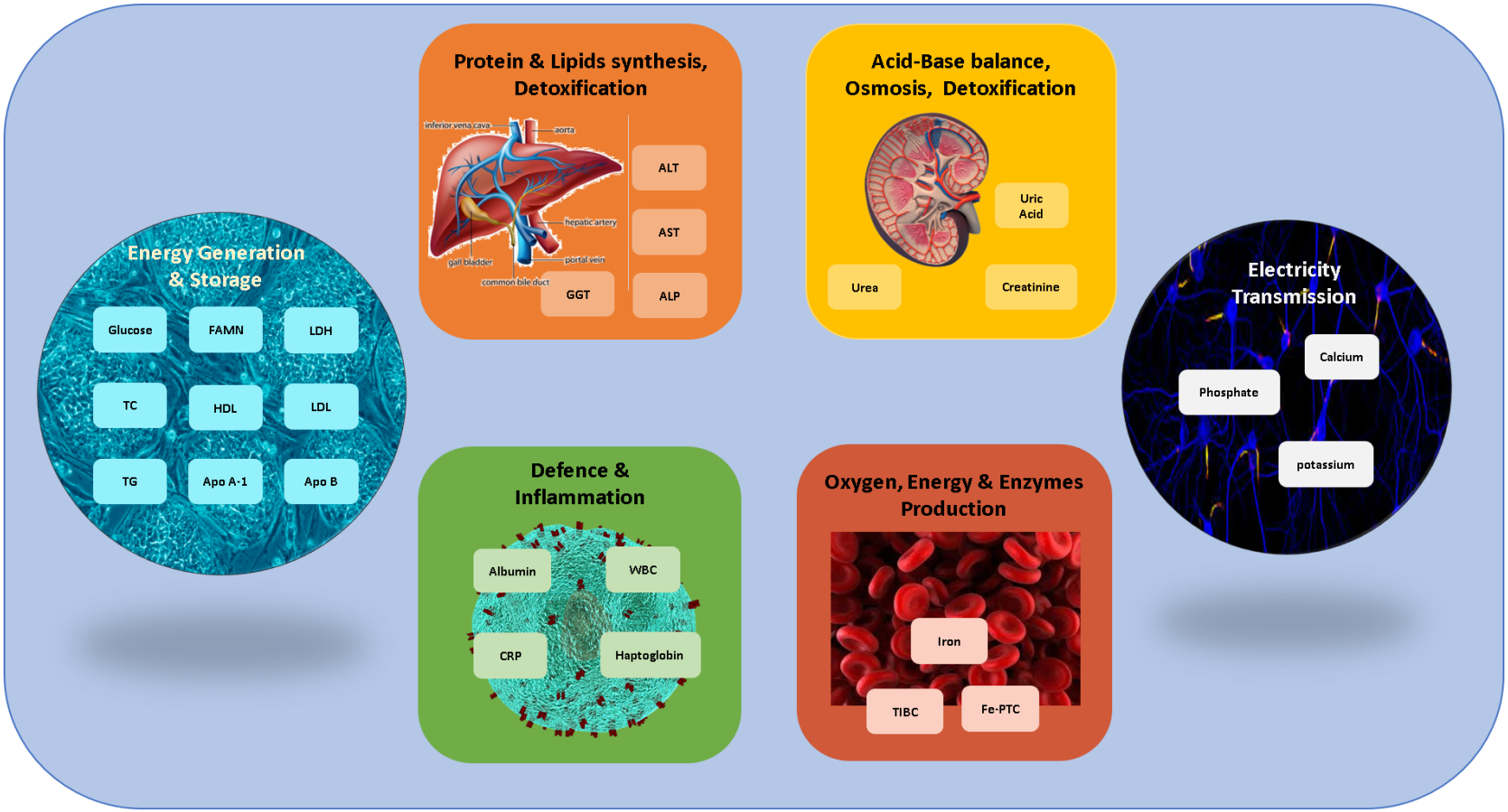
Table 3. Fully automated laboratory methods with automatic calibration were performed at one accredited laboratory (CALAB) to measure the serum biomarkers examined in this study.

Markers	Method	Total imprecision*
Glucose	GOD-PAP method: Enzymatic colorimetric test in which glucose in serum reacts with oxygen to give gluconate and hydrogen peroxide in the presence of glucose oxidase.	<2.2% CV
Fructosamine	This technique is based on the reducing ability of fructosamine in an alkaline solution.	≤5.0% CV
Total cholesterol	CHOD-PAP: Enzymatic cholesterol assay based on cholesterol esterase and cholesterol oxidase conversion followed by a Trinder-type sequence of reactions.	≤2.7% CV
Triglycerides	GPO-PAP: Enzymatic determination of glycerol with glycerol-phosphate-oxidase (GPO) after hydrolysis with lipoprotein lipase.	≤5.0% CV
Apolipoprotein A-I	Immunoturbidimetry using polyclonal antisera from Orion (Helsinki, Finland).	<4.0% CV
Apolipoprotein B	Immunoturbidimetry using polyclonal antisera from Orion (Helsinki, Finland).	<4.0% CV
Lactate dehydrogenase	LD (EC 1.1.1.27) in serum catalyses the reaction in which pyruvate is reduced to L-Lactate by dihydronicotinamide adenine dinucleotide ($NADH_2$ causes a decrease in absorbance at 340nm with resepect to time. Within the range of the method, the rate of decrease of absorbance, measure at 340nm, is directly proportional to the LD activity).	<4% CV
Alanine aminotransferase	Enzymatic UV-test according to International Federation for Clinical Chemistry (including incubation with pyridoxal phosphate).	≤6.0% CV
Aspartate aminotransferase	Enzymatic UV-test according to International Federation for Clinical Chemistry (including incubation with pyridoxal phosphate).	≤6.0% CV
Gamma-glutamyl transferase	Enzymatic colorimetric test using reagents from Randox Laboratories Ltd, Crumlin, UK.	≤6.0% CV
Alkaline Phosphatase	AP (EC 3.1.31) in serum dissociates parnitrophenyl phosphate into paranitrophenol and phosphate. The formation of para-nitrophenol causes an increase in absorbance with respect to time. Within the range of the method, the rate of increase of absorbance, measured at 405nm, is directly proportional to the ALP activity.	<5% CV at 1.5 (Microkat//L) <4% between 3.3 and 7.5 (Microkat/L)
Creatinine	Jaffe method (kinetics) with deproteinization. Creatinine in serum reacts with alkaline picrate to form a colour complex. The formation of this complex causes an increase in absorbance with respect to time. Within the range of the method, the difference in absorbance at 490nm is directly proportional to the creatinine concentration.	<3% CV
Albumin	Bromocresol green method.	≤2.0% CV
Leukocytes	The STKS is a haematology flow cytometer that automatically performs blood cell counting from whole blood samples.	<2.7% CV

C-reactive protein	Immunoturbidimetry using polyclonal antisera from Orion (Helsinki, Finland).	≤12.0% CV
Haptoglobin	Immunotubidemetric determination of haptoglobin. Reagents from DAKO A/S, Glostrub, Denmark.	≤5.0% CV
Iron	Acidification with citric acid in order to dissociate the Fe3 transferring complex.	<5% CV
Total Iron binding capacity	Fe^{3+} is added to the serum, in excess of that capable of being bound by the protein transferrin. By the addition of 2, 4, 6 –tri(2-pyridyl) – 1, 3, 5 -triazine, a coloured complex is formed with the iron which is not bound. The absorbance of the complex is measured. The difference between the quantity of iron added to serum and that found to be in excess is equal to the unsaturated iron-binding capacity of the serum.	≤5.0% CV
Phosphate	Formation of the phosphomolybdic acid complex.	<4% CV
Calcium	A colorimetric method was used for the measurement of total serum calcium.	<2.5% CV

* CV: Coefficient of Variation

Figure 20. Panel of biomarkers studied in this chapter. The biomarkers are displayed characterising different biological processes involved which represent main metabolic pathways in which the set of routinely collected serum biomarkers may play a role. These metabolic pathways are fundamental for the body homeostasis and as a consequence they might also be relevant in carcinogenesis.



*The following abbreviations have been used in the illustration: Fructosamine (FAMN), Lactate dehydrogenase (LDH), Total Cholesterol (TC), High-density lipoproteins (HDL), Low-density lipoproteins (LDL), Triglycerides (TG), Apoliprotein A-1 (Apo A-1), Apoliprotein B (ApoB), Alanine Amino Transferase (ALT), Aspartate Amino Transferase (AST), Gamma-Glutamyl Transferase (GGT), Alkaline Phosphatase (ALP), Total Iron Binding Capacity (TIBC), Transferrin Saturations (Fe-PCT), Leukocytes (WBC) and C-Reactive Protein (CRP).

A brief description of each of the pathways included in the projects, explaining the specific markers and their clinical application is explained below:

Energy metabolism

The generation of energy in human cells is mainly driven by the glycolysis cycle, a metabolic cycle that transforms glucose into pyruvate to generate the fuel molecules of the body, adenosine triphosphate (ATP) and reduced nicotinamide adenine dinucleotide (NADH) (189). The pyruvate resulted from the glycolysis is then used to synthesize fatty acids. Fatty acids produce the higher fraction of ATP of all molecules when oxidized by beta oxidation and citric acid cycle, so therefore fatty acids are the main storage form of fuel in humans. Fatty acids also are key parts of the cell membranes on the form of phospholipids (190). Glycolysis is an oxygen independent pathway, named as one of the hallmarks of cancer given that cancer cells reprogram the energy metabolism limiting their energy metabolism largely to glycolysis (3, 4). These two main energy metabolism cycles have also been targeted by Pavlova as hallmark of cancer pathways: the deregulated uptake of glucose and amino acids, the use of opportunistic modes of nutrient acquisition as lipids and the use of glycolysis/TCA cycle intermediates for biosynthesis and NADPH production (29).

From these two main energy producer cycles of the body, eight different established screening markers have been considered in my analysis: Glucose, Fructosamine (FAMN), Total Cholesterol (TC), High-density lipoproteins (HDL), Low-density lipoproteins (LDL), Triglycerides (TG), Apoliprotein A-1 (Apo A-1) and Apoliprotein B (ApoB). Lactate dehydrogenase (LDH) is another marker involved in the energy cycle and consequently, it was also considered in this section.

The following section explains the utility of these screening markers from the current clinical perspective. The fasting glucose blood test measures current glucose levels in blood and it is performed for diagnosis of a pre-diabetes condition. Individuals with impaired fasting glycaemia or abnormal elevated blood glucose

levels in the body after fasting, are individuals that are not able to process glucose efficiently and consequently at higher risk of type 2 diabetes, cardiovascular conditions or metabolic syndrome (191). Fructosamine (FAMN), a compound result of the combination of sugar and an amine, is also employed in diabetes diagnosis. Fructosamine blood test determines the amount of total serum proteins that have undergone glycation which reflects the average of 1-3 weeks glucose levels in blood. For diabetes management, the use of the haemoglobin A1c test, which measure the glycation haemoglobin levels, is preferred to FAMN, because it reflects the glucose levels in blood over a three months period, however this biomarker was not available in CALAB sample set (192).

Cholesterol is a fatty acid, essential component of cell membranes and a precursor of hormones. Cholesterol is transported through the blood stream in lipoproteins and when there is an excess of cholesterol in the blood stream, it can be deposited in plaques in the blood vessels which can cause future block of the vessels, which increases the risk of cardiovascular events. Total cholesterol blood test determines all the lipid components in the blood stream and is used alone or combine with the other lipids, in a lipid panel to establish the cardiovascular and heart health status of an individual. High-density lipoproteins (HDL) and low-density lipoproteins (LDL) are part of the lipid profile blood test. HDL is one type of lipoproteins that carries cholesterol and its function is to remove excess of cholesterol from the blood and carry it to the liver for disposal. In contrast, LDL, another type of lipoprotein that carries cholesterol, tends to deposit the excess of cholesterol in the wall of the blood vessels.

Triglycerides (TG), also a major component of the lipid profile, are normally storage in adipose tissue, but between meals are released to blood as an energy source for the body (193). The standard lipid profile will include all the above-mentioned components to assess the lipid levels in blood for diagnosis of cardiovascular disease.

Apoliprotein A-1 (Apo A-1), is the major component of the HDL molecules in plasma, accept for fats from within cells for transport, while Apoliprotein B (Apo B) is a component mainly found in LDL. These molecules are not standard health biomarkers, however the ratio apo B/apo A-1 has been found to have a stronger correlation with cardiovascular disease than the conventional lipid profile (186, 187, 194). Consequently, these two biomarkers have also been included in project B.

Lactate dehydrogenase (LHD) is an enzyme that catalyses the reaction from lactate to pyruvate, the last step in the anaerobic glycolysis. The enzyme appears extensively in body tissues, as blood, muscles, brain or kidney, and it is released into the blood stream when a tissue damage exits. Consequently, it is used as a marker of tissue damage events such as heart failure, anaemia, pancreatitis, liver or lung disease (195). Given its key role in the glycolysis, and the increase of glycolysis cycle in cancer cells, raised LHD levels in blood can be used as an unspecific tumour marker (196).

Protein, lipid synthesis and detoxification in liver

The liver has a key role in maintaining the internal body homeostasis. It is responsible for the synthesis of multiple proteins such as clotting factors and albumin, assembles carbohydrates, cholesterol and triglycerides, and synthesises glycogen to be stored in muscle and liver, together with bile production for food digestion. Moreover, the liver stores vitamins and iron. It filters the blood coming from the digestive system and detoxifies chemicals and metabolises drugs as alcohol, transforming ammonia into urea that is excreted to urine by the kidneys (197).

Four different established screening markers for liver function have been considered in my analysis: Alanine Amino Transferase (ALT), Aspartate Amino Transferase (AST), Gamma-Glutamyl Transferase (GGT) and Alkaline Phosphatase (ALP). ALT is an aminotransferase enzyme commonly found in most tissues, mainly in liver and kidney, and fluids, apart from urine. ALT catalyses the transfer of the

amino group from glutamate to pyruvate to form ketoglutarate and alanine, it is an important enzyme in the gluconeogenesis cycle. The ALT blood test is a sensitive marker for the diagnostic evaluation of liver health status. It has also been appointed as a good marker of health specifically for cardiovascular disease and metabolic syndrome. The ALT test measures levels of the enzyme in plasma that will be elevated if the liver is damaged or inflamed (198).

Aspartate Amino Transferase (AST), another aminotransferase, is always measured in conjunction with ALT for liver disease diagnosis, given its faster but shorter release into the blood stream when the liver is damaged. AST catalyses the transfer of an amino group between aspartate and glutamate to form oxaloacetate and glutamate, being key on the amino acid metabolism. AST is found in multiple organs as liver, heart, muscle, red blood cells because of this AST test considered less specific that ALT (199).

Gamma-Glutamyl Transferase (GGT) is a transferase enzyme that catalyses transfer of gamma-glutamyl groups, being an important part of the drug and xenobiotic detoxification pathway. GGT is found in the cell membrane of many tissues, mainly in liver. GGT blood test is performed predominantly as a marker of liver disease, but also of the biliary system and pancreas (200). Recently is also been used for CDV diagnosis (201).

Alkaline Phosphatase (ALP) is a phosphatase enzyme that dephosphorylates compounds. This enzyme plays a main role in liver function and skeleton and hence the ALP blood test is mainly used for diagnosis of liver diseases such as hepatitis and bone disorders, but also biliary system and kidney (202).

Acid-Base balance, osmosis and detoxification in kidney

The kidneys have a major role maintaining the homeostasis in the human body by preserving main functions such as the acid-base balance, extracellular fluid volume, blood pressure, and the concentration of electrolytes and the removal of toxins.

These functions are processed by filtering, reabsorbing, secreting and excreting the blood plasma components received through the renal arteries. Some substances such as glucose, amino acids are reabsorbed, while other substances such as ammonium and uric acid are excreted to the urine via the bladder (203).

Three different established screening markers for kidney health status have been considered in my analysis: Creatinine, Urea and Uric acid (Urat).

Creatinine, is an excreted product from the muscle metabolism. It is produced in the liver and then transported to other organs to be transformed in a high-energy compound. When it is excreted to kidneys it can hardly be reabsorbed, making an ideal candidate to assess kidney function. Serum creatine levels are correlated with the glomerular filtration rate which is used to diagnose kidney disease, consequently high levels of serum creatinine may suggest a kidney malfunction (204).

Urea nitrogen is an excreted product from protein catabolism in the liver. Urea is excreted to the blood stream and it is mainly filtered by the kidneys to the urine. High levels of urea nitrogen in the blood stream may suggest abnormal kidney function (205).

Uric acid (Urat) is a product of the catabolism of the nucleotides, which is normally released to blood, filtered by the kidneys and excreted to urine (206). High or low levels of uric acid in blood can be associated with health conditions such as cancer, cardiovascular disease or gout (high uric acid levels) and kidney or liver disease (low uric acid levels) (207, 208).

Defence and inflammation in the immune system

The immune system is a body system that protects the organism against disease through sophisticated mechanisms. The immune system in a simple definition is able to detect external agents or pathogens such as bacteria, viruses or other

external organisms and neutralise them. The different mechanisms that allow the human body to defend infections can be classified into two main categories: the innate or humoral immunity and the adaptive or cell-mediated immunity.

Processes part of the innate immune system are the surface barriers, such as mechanical, chemical and biological barriers against invasion of external agents, the inflammatory processes that are the first response to infection as for example the raise of body temperature, the leukocytes (WBC) that recognise pathogens and neutralise them, and finally the complement system that works as a biochemical reaction that attacks the surface of the exogenous cells. The adaptive immune system consists of the development of an immune memory that remembers and recognises pathogens. This mechanism is performed by the memory cells or lymphocytes, a special type of leukocytes (209).

If there is a malfunction of the immune system, it can result in autoimmune disease, inflammatory disease, immunodeficiency disease and cancer. Due to the major role of the immune system, evading immune response and inflammation have been considered main hallmarks of cancer (3, 4).

Four different established screening markers for the immune health status have been considered in my analysis: Albumin, White Blood Cells (WBC), C reactive protein (CRP) and Haptoglobin.

Serum albumin is the main plasma protein produced by the liver and its main function is to maintain fluid within the circulation system. It also carries low density proteins in blood. Abnormal serum albumin levels can be associated with liver or kidney disease (210). Moreover, due to capacity of binding with polysaccharides and other bacterial products, it can modulated the inflammatory reaction (211).

Leukocytes or white blood cells (WBC) are the main body immune cells that protect against infections and pathogens. WBC are a diverse family of cells including the

macrophages, neutrophils, dendritic cells, innate lymphoid cells, mast cells, eosinophils, basophils, and natural killer cells. WBC can be found in blood and lymphatic system and throughout the body. The WBC count in blood is a standard clinical marker of the body health status, indicating disease when numbers are higher or lower that the normal range.

C - reactive protein (CRP) is a plasma protein synthesised in the liver that is released into the blood in response to inflammation. CRP blood tests are usually performed to diagnose inflammation, but it has also been associated with cancer, CVD and metabolic syndrome (212, 213).

Haptoglobin is a plasma protein synthesised in the liver that binds free haemoglobin, the complex is soon degraded and the iron is recycled. Haemoglobin is carried by red blood cells so when these cells are damaged, haemoglobin is released in blood and captured by haptoglobin. In conditions as haemolytic anaemia, haptoglobin serum levels decrease. Haptoglobin serum levels are also associated with liver disease, inflammatory diseases, including infections, atherosclerosis, and autoimmune disorders (214).

Oxygen, energy & enzymes production in the iron metabolism

Iron metabolism is responsible for essential body functions such as haematopoiesis to produce haemoglobin which transports oxygen in red blood cells, it generates energy at the mitochondria through the cellular respiration, produces enzymes and hormones and promotes the immune system by depriving of iron load to bacteria which results in bacteria growth decrease. Control of iron homeostasis is essential for the body given that both iron overload and iron deficiency are causes of important human diseases, such as hemochromatosis or anaemia leading to tissue hypoxia; body iron unbalance has been associated with major diseases as CVD and cancer (215-220). Moreover, free iron is extremely cytotoxic, therefore iron is always found bound to proteins and cofactors, as the heme group in haemoglobin. A small amount of the iron intake from diet is absorbed in the intestines and then

released into plasma as an iron-transferrin complex which can be transferred to the liver or spleen to be stored as iron-ferritin complex or to the bone narrow to create a heme group part of haemoglobin carried by the red blood cells (221, 222).

Three different established screening markers for the iron body balance have been considered in my analysis: Serum Iron, Total Iron Binding Capacity (TIBC) and Transferrin Saturations (Fe-PCT).

These three biomarkers are closely interrelated to comprehend the amount of iron in plasma and are usually combined to report iron deficiency in the body. Serum Iron blood test measures the iron that is in circulation bound to transferrin protein and to assesses the extent (%) of iron transport by transferrin the Transferrin Saturations (Fe-PCT) marker is used. Total Iron Binding Capacity (TIBC) determines the capacity of blood to bind iron with transferrin (223).

Electricity transmission with electrolytes

Electrolytes are chemicals that have the capacity to conduct electricity when dissolved in water, such as phosphate, calcium, potassium, sodium or magnesium, and are essential for multiple body functions. Electrolytes regulate the muscle and nerve function, regulate body acidity and body hydration and help to heal damage tissue. Therefore, electrolytes homeostasis needs to be tightly balanced in the body. Excess or deficiency of electrolytes can cause deficiency in muscle function and nerve transmission and it has been associated with different diseases as cancer (224). Only a small percent of the electrolytes is normally present in blood.

Three different electrolytes have been considered in my analysis: Phosphate, Calcium and Potassium.

Phosphate is a chemical closely related to bone growth, muscle and nerve function but also to maintain acid-base balance. A phosphate blood test is normally

performed in conjunction with calcium to detect kidney or gastrointestinal alterations (225).

Calcium is a chemical mainly stored in the bone and together with phosphorous builds and maintains bone. It has other important functions such as transmission of the cell signalling for the correct functioning of muscles, nerves, and the heart. Total calcium is normally measured to diagnose alterations related to bones, heart, nerves and kidneys (226).

Potassium is a chemical essential to maintain the muscle function transmitting signal between muscles and nerves, especially important for heart function. It also transports nutrients to cells and helps to eliminate toxins from cells. Serum potassium is usually assessed to diagnose kidney and heart disease (227).

All the above-mentioned biomarkers have been investigated in the two projects discussed in this chapter. It would have been desirable to include more markers covering other pathways in the analyses, however the study was limited by the availability of markers in the CALAB database and the number of blood results available simultaneously at baseline. However, all markers were carefully chosen to be as representative as possible of important metabolic pathways.

III. Project A: The blood exposome in AMORIS

The findings of this project are currently under review with Plos One journal.

a. Rationale

Chronic diseases present a multi-factorial aetiology with an intricate gene-environment interplay (228). However, classical observational epidemiological studies have mainly studied single factors in relations with health outcomes, including complex diseases such as cancer or cardiovascular disease (229, 230). From a mechanistic point of view, this approach has been fundamental to understand the role of those factors in specific biological pathways and their efficacy as diagnostic markers in different diseases (231, 232). However, given the multifactorial nature of complex diseases, where multiple pathways at multiple levels of complexity are dysregulated during pathogenesis, a systemic approach is necessary to decode the biological mechanisms that drive disease (3, 4, 29).

With the advent of genomics, there has been increasing interest in trying to unravel the environmental component of disease, especially since the introduction of the theoretical concept of the exposome by Professor Christopher Wild in 2005 (106). Multiple efforts towards the practical implementation of the exposome have been made since (87, 107, 109, 114, 233-240).

Based on a set of routinely collected blood biomarkers available in the Swedish AMORIS database, the current project is set out to use an exposome-based approach. More specifically, given the top down approach presented by Rappaport in 2014 to explore blood metabolites, carriers of both internal and external exposure components (117), I aimed to investigate the reciprocity between clinically meaningful blood biomarkers and their association with external factors in an attempt to evaluate a subset of the internal-external exposome components in a given population. The project is not a full application of the exposome, but uses a similar framework and aims to evaluate associations between internal and external components of the exposome.

Thus, to explore the synergy in 21 clinically meaningful blood biomarkers and external factors, this study aimed to (1) evaluate interactions in commonly assessed health biomarkers via their interplay into novel groups and (2) assess how the external socio-demographic factors influence these biomarkers.

b. Methods

i. Study population

This study utilises individuals from the Apolipoprotein Mortality-related Risk (AMORIS) database, as described above. For this project, I included all individuals aged 20 or above with baseline measurements of 21 standard of care metabolites (n=154,207), who did not develop a cancer diagnosis within the next three years. From the CALAB database, information on age, sex and fasting status was obtained. Information on socio-economic status and education levels was obtained from national census data (241). In the database, socio-economic status was coded as blue or white collar. The panel of biomarkers included in the analysis were: Glucose, Fructosamine (FAMN), Total Cholesterol (TC), Triglycerides (TG), Lactate Dehydrogenase (LDH), Alanine Amino Transferase (ALT), Aspartate Amino Transferase (AST), Gamma-Glutamyl Transferase (GGT), Alkaline Phosphatase (ALP), Creatinine, Urea, Uric acid (Urat), Albumin, C-Reactive Protein (CRP), Haptoglobin, Iron, Total Iron Binding Capacity (TIBC), Transferrin Saturations (Fe-PCT), Phosphate, Calcium and Potassium. All measurements were measured on the same day and the biomarkers were selected to represent the main metabolic pathways (see above). During the time of laboratory measurements, high-sensitivity CRP was not available. Therefore, CRP levels <10mg/L were not measurable.

No specific outcome was defined in this analysis as the project aimed to understand the interrelations between all the different markers (internal and external).

ii. Statistical Analysis

The statistical pipeline in this project is divided into four main exploratory analyses to understand the interactions between the different serum markers and potential external factors: (1) correlations to understand whether there is collinearity

between the markers, (2) hierarchical clustering to assess whether there are homogeneous subgroups in the dataset, (3) principal component analysis (PCA) to summarise if there is a set with a smaller number of representative variables that collectively explain most of the variability in the original set, and (4) multivariate analysis of variance (MANOVA) to assess the interaction between internal and external markers in the dataset. This approach has been used previously to investigate metabolic markers of exposure in the study by *Chadeau-Hyam et al*. (11, 129, 242-244).

The normal distribution was explored for all the markers included in the analysis and given the fact that many variables were not normally distributed, a normalisation was applied to all variables prior to analysis, by transforming all the values to logarithmic values. Age was categorised into four groups: <40, 40<50, 50<=65 and >65. The 50-65 group was included to consider menopausal status in females. The last category, included all the individuals that were retired and therefore may be different to the rest of the population strata studied.

The four explanatory analyses are explained in more detail below:

(1) Correlations between all markers were calculated using Spearman's rank order correlations to produce a default 21 x 21 data frame matrix between all pairs of variables that explores the interplay between variables. This same matrix was calculated for each category of the external markers, sex, age, SES and education.

R package *corrplot* (245) was employed to visualise the matrixes and using *iGraph* R package (246), a network analysis (cut-off >0.35 for branches) was performed as a sensitivity analysis to check the interaction between markers.

(2) Hierarchical clustering was performed in standardised data using Z transformation (mean =0, standard deviation = 1) applying the function *hclust* of the R package *stats* to explore subgroups of markers within the

data. Hierarchical clustering, network analysis and correlations studied the internal interactions between the markers, consequently no further analysis exploring interactions were conducted.

(3) Prior to the analysis, a Z transformation of the dataset was required for **Principal component analysis**. R package *principal* was applied to perform the PCA to reduce the dimension of the data into few components that explain maximum variation of the dataset (247). A scree-plot graph was produced to visualise the variance of the first 10 components (default number of components displayed by *principal*), moreover the proportion of variance explained (PVE) in each component and the cumulative proportion were performed to investigate the loadings (weight on the total dataset) of these components.

(4) **Multivariate analysis of variance** was calculated to assess the statistical difference in mean values for each blood marker and each of the external socio-demographic factors in SAS. Levene's t-test was used to assess the homogeneity of the variance of the markers. Welch's t-test was then used in case of equal means but unequal variances, while Tukey's test was applied to assess pairwise comparison of means when there were more than two categories ($p<0.05$).

Data management was performed using SAS version 9.4 (SAS Institute, Cary, NC, USA) and data analysis was conducted with R version 2.13.2 and R studio 3.3.2 (R Foundation for Statistical Computing, Wien, Austria) (248, 249).

c. Results

The tables and figures resulted from the analyses are displayed at the end of this section to facilitate the flow of information.

Descriptive statistics of the baseline characteristics of study population is displayed in Table 4. A total of 154,207 participants were included in the study (53% male, 47% female). More than 90% of the population were “gainfully employed” and consequently, in the strata between the 40 to the 65-age group. More people had “white collar jobs” (47.54%) as compared to “blue collar jobs” (42.69%). 45% of the population had middle level of education status.

Spearman correlation’s results are displayed in Figure 21, which presents a 21 by 21 scatter plot matrix in which the biomarkers are plotted against each other. The pairwise correlations were considered significant at a p-value < 0.01. The stronger correlations found in the dataset were: Serum Iron and Fe-pct. (r=0.93), part of the iron metabolism and the liver enzymes ALT and AST (r=0.68) and ALT and GGT (r= 0.58). Other correlations were present within markers of similar pathways with r values around 0.4. Moreover, a positive association was observed between AST and LHD (r=0.44) and between Calcium and Albumin (r=0.42). Furthermore, I explored the correlations between the internal markers by categories of external factors (Figures 22 to 25). A strong correlation was observed within markers of related pathways for the employed or missing category of the socio-economics status. Interestingly, for increasing categories of education (low, medium, high), significant levels of correlations were observed for Albumin vs Calcium: r=0.39, r=0.43, r=0.45, respectively and Creatinine vs Uric Acid: r=0.43, r=0.46, r=0.48. The magnitude of correlation between Glucose and Fructosamine decreased with higher education (r=0.50, r= 0.42, r=0.37, respectively) (Figure 23). Trends were observed with increasing age from 21 years up to 60 years of age (figure 24). Stronger positive correlations between markers were observed with females compared to males (e.g., between liver and lipid markers and ALT, LDH and Urat) (figure 25).

The hierarchal clustering dendrogram is displayed in Figure 26. A cut-off at a height of 580 points that produced three clusters was selected to facilitate the visualisation of the clusters, presenting markers with similar interactions and metabolic pathways and allowing comparisons between groups, however the full dendrogram

is explained below. The first cluster contained only Iron and FE-pct, which is was consistent with the strongest correlation observed in the dataset. The second group contained TIBC, Albumin, Calcium, Phosphate and Potassium. The third included the rest of the markers in subgroups representing connected pathways, LDH, ALP, GT, ALT and AST clustered together, one subgroup contains Urea, Creatinine and Uric Acid, another one includes Glucose, Fructosamine, TC and Triglycerides and finally CRP and Haptoglobin grouping together. Therefore, the results were consistent with the output of the correlation analysis.

The principal component analysis identified 21 components that explained the variance in the dataset. Figure 27 displays the scree plot of the PCA analysis for the variances of the first 10 out of 21 principal components, a default-plot of the R package *principal*, with its associated proportion of variance explained and cumulative proportions. Out of 21 principal components, the first 12 had Eigen values greater than 1. The first component had an Eigen value > 3 and accounted for 18% of the variance in the data. The proportion of the first 10 components accounted for 74% of the variance data. The first 12 principal components accounted for 80% of the data. The contents of these components were sparse and did not showed any specific pattern. The first component was enriched on markers of liver metabolism, lipid, glucose and kidney functioning, while the second component showed few markers including iron and liver markers with negative and positive values. The third component presented the same markers as the first component, with half of them with negative values (Figure 28). Because the proportions of variance were distributed throughout all 21 components, with the first four components explaining 43% of the data, no further analysis was carried out using principal component analysis.

The results of the multivariable analysis of variance are illustrated in Table 5. Statistical differences in mean were observed for all variables between “blue collar” vs “white collar”. A similar trend occurred for males vs females, except for TC where no difference between the means was observed. In general, the serum biomarkers

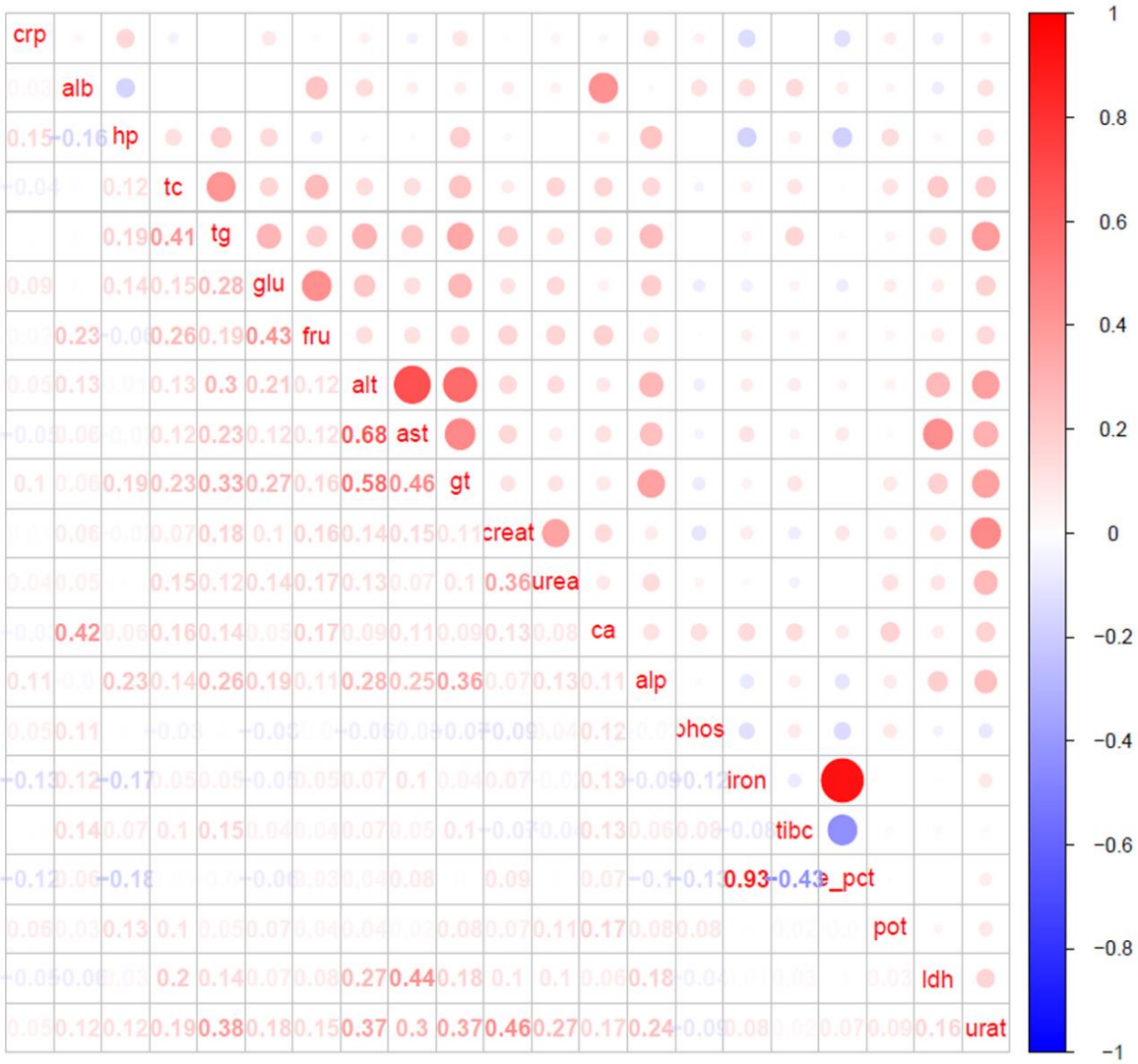
were higher in the lower education group, males and older age, respectively. For the class variable education, no differences between means were observed between medium levels of education and high levels of educations for markers CRP, Phosphate and Iron. Age categories showed statistical differences in mean for most of the variables between the 4 categories, except from Creatinine and ALP for ages <40 – 50 and GGT, Calcium and Phosphate for ages 50 to >65.

Table 4. Descriptive statistics of the baseline characteristics of the study population.

	Total N = 154,207 (100%)
Age (years)	
Mean (SD)	47.31 (13.77)
Median (q1, q2)	47.07 (37.76, 58.27)
<40 (%)	46661 (30.26)
40-50 (%)	42986 (27.88)
50-65 (%)	34588 (22.43)
>65 (%)	29972 (19.44)
Sex	
Male (%)	81796 (53.04)
Female (%)	72411 (46.96)
Socio Economic Status	
White Collar (%)	67763 (43.94)
Blue Collar (%)	75747 (49.12)
Not gainfully employed or missing (%)	10697 (6.94)
Education Status	
Low (%)	42630 (27.64)
Middle (%)	68580 (44.47)
High (%)	42997 (27.88)
Biomarkers	**Mean (q1, q2)**
C-Reactive protein (mg/L)	5.33 (1.00, 6.00)
Albumin (g/L)	42.80 (41.00, 45.00)
Haptoglobin (g/L)	1.05 (0.90, 1.20)
Cholesterol (mmol/L)	5.63 (4.80, 6.30)
Triglycerides (mmol/L)	1.35 (0.70, 1.60)
Alanine amino transferase (IU/L)	0.44 (0.24, 0.50)
Aspartate amino transferase (IU/L)	0.39 (0.28, 0.42)
Gamma-glutamyl transferase (IU/L)	0.48 (0.23, 0.50)
Glucose (mmol/L)	4.99 (4.40, 5.20)
Fructosamine (mmol/L)	2.10 (1.95, 2.21)
Creatinine (µmol/L)	80.80 (72.00, 89.00)
Urea (mmol/l)	5.05 (4.10, 5.80)
Calcium (mmol/L)	2.39 (2.32, 2.45)
Alkaline phosphatase (IU/L)	2.64 (2.00, 3.10)
Phosphate (mmol/L)	1.05 (0.90, 1.10)
Iron (µmol/L)	17.70 (14.00, 21.00)
Total iron binding capacity (µmol/L)	59.60 (54.00, 64.00)
Transferrin Saturations (µmol/L)	0.30 (0.23, 0.36)
Potassium (mmol/l)	4.22 (4.00, 4.40)

Lactate dehydrogenase (µmol/L)	5.86 (5.10, 6.40)
Uric acid (µmol/L)	289.00 (236.00, 334.00)

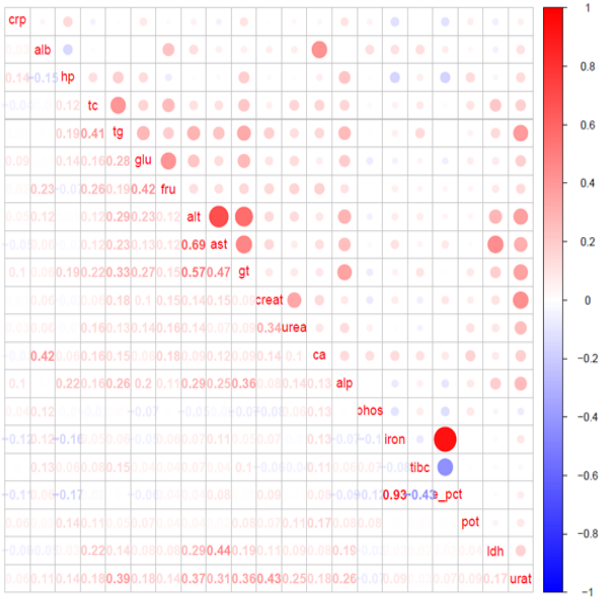
Figure 21. Spearman's rank-order Correlation matrix between all 21 blood markers is displayed in the 21 by 21 scatter plot matrix in which the biomarkers are plotted against each other. The strength of the correlation is represented by the size of the circles (bigger higher r value) and the sign of the correlation is displayed using the red and blue palette (red means positive correlation while blue represents negative correlations). The actual r values are also displayed in the plot.



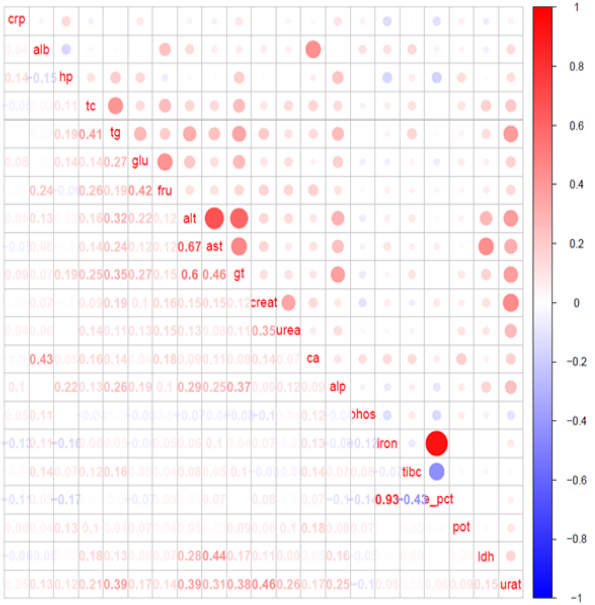
*The following abbreviations have been used: crp = c-reactive protein, alb = albumin, hp = haptoglobin, tc = total cholesterol, tg = triglycerides, glu = glucose, fru = fructosamine, alt = alanine amino transferase, ast = aspartame amino transferase, gt = gamma glutamyl transferase, creat = creatinine, ca = calcium, alp = alkaline phosphatase, phos = phosphate, tibc = total iron binding capacity, fe-ptc = transferrin saturations, pot = potassium ldh = lactate dehydrogenase and urat = uric acid.

Figure 22. Scatter plot matrix's using spearman's rank correlations for markers: 21 by 21 scatter plot matrix in which the biomarkers are plotted against each other by categories of socio-economics status. The strength of the correlation is represented by the size of the circles (bigger = higher r value) and the sign of the correlation is displayed using the red and blue colours (red means positive correlation while blue represents negative correlations). The actual r values are also displayed in the plot.

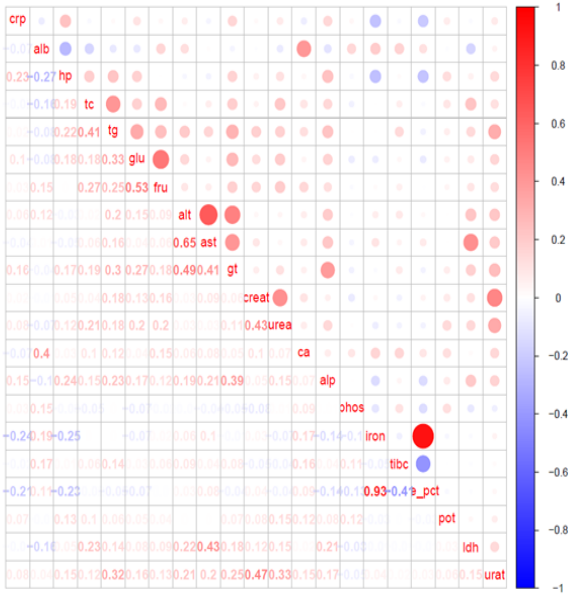
Blue Collar



White Collar

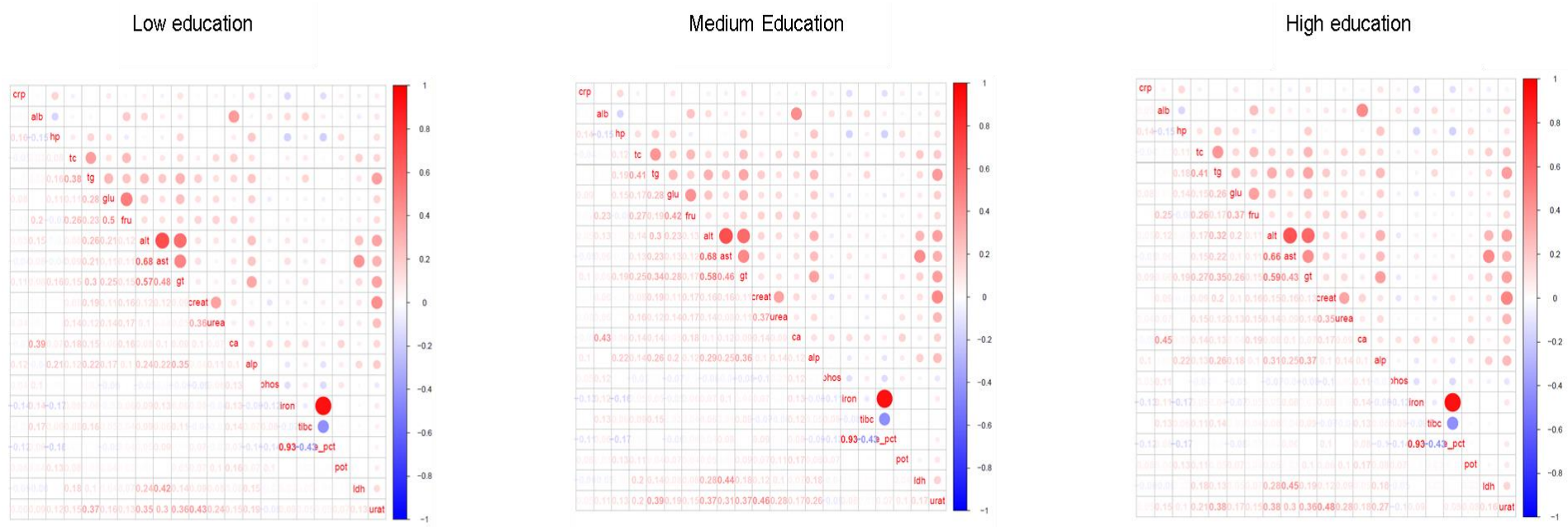


Unemployed/missing



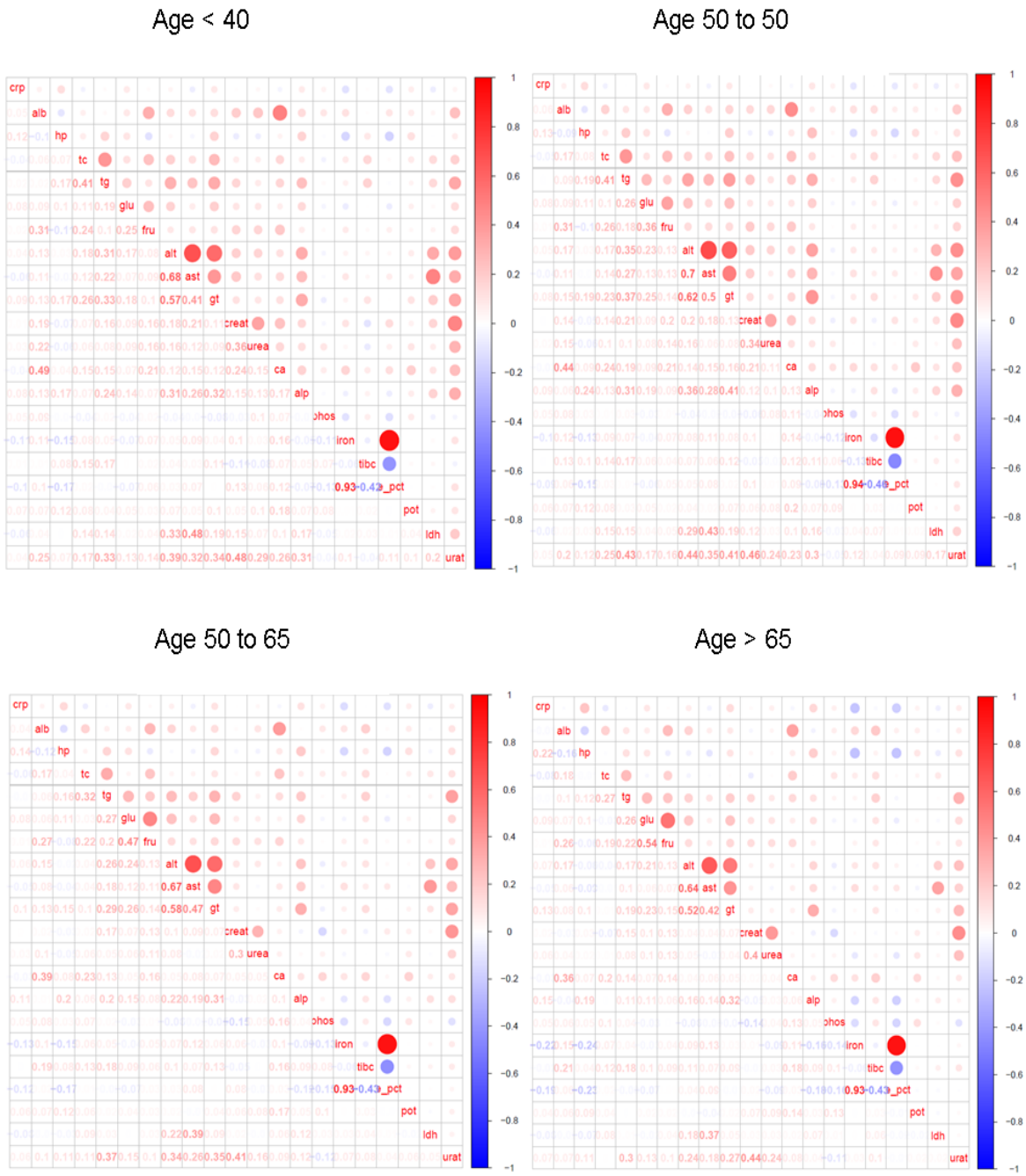
*The following abbreviations have been used: crp = c-reactive protein, alb = albumin, hp = haptoglobin, tc = total cholesterol, tg = triglycerides, glu = glucose, fru = fructosamine, alt = alanine amino transferase, ast = aspartame amino transferase, gt = gamma glutamyl transferase, creat = creatinine, ca = calcium, alp = alkaline phosphatase, phos = phosphate, tibc = total iron binding capacity, fe-ptc = transferrin saturations, pot = potassium ldh = lactate dehydrogenase and urat = uric acid.

Figure 23. Scatter plot matrix's using spearman's rank correlations for markers: the 21 by 21 scatter plot matrix in which the biomarkers are plotted against each other by categories of education status. The strength of the correlation is represented by the size of the circles (bigger higher = r value) and the sign of the correlation is displayed using the red and blue colours (red means positive correlation while blue represents negative correlations). The actual r values are also displayed in the plot.



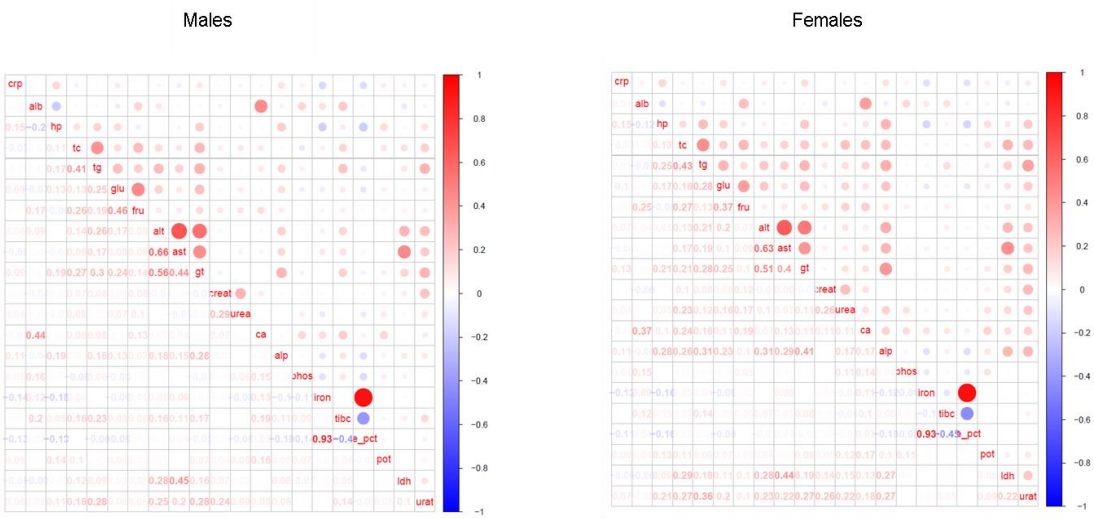
*The following abbreviations have been used: crp = c-reactive protein, alb = albumin, hp = haptoglobin, tc = total cholesterol, tg = triglycerides, glu = glucose, fru = fructosamine, alt = alanine amino transferase, ast = aspartame amino transferase, gt = gamma glutamyl transferase, creat = creatinine, ca = calcium, alp = alkaline phosphatase, phos = phosphate, tibc = total iron binding capacity, fe-ptc = transferrin saturations, pot = potassium ldh = lactate dehydrogenase and urat = uric acid.

Figure 24. Scatter plot matrix's using spearman's rank correlations for markers: 21 by 21 scatter plot matrix in which the biomarkers are plotted against each other by categories of age. The strength of the correlation is represented by the size of the circles (bigger higher = r value) and the sign of the correlation is displayed using the red and blue colours (red means positive correlation while blue represents negative correlations). The actual r values are also displayed in the plot.



*The following abbreviations have been used: crp = c-reactive protein, alb = albumin, hp = haptoglobin, tc = total cholesterol, tg = triglycerides, glu = glucose, fru = fructosamine, alt = alanine amino transferase, ast = aspartame amino transferase, gt = gamma glutamyl transferase, creat = creatinine, ca = calcium, alp = alkaline phosphatase, phos = phosphate, tibc = total iron binding capacity, fe-ptc = transferrin saturations, pot = potassium ldh = lactate dehydrogenase and urat = uric acid.

Figure 25. Scatter plot matrix's using spearman's rank correlations for markers: 21 by 21 scatter plot matrix in which the biomarkers are plotted against each other by categories of gender. The strength of the correlation is represented by the size of the circles (bigger higher = r value) and the sign of the correlation is displayed using the red and blue colours (red means positive correlation while blue represents negative correlations). The actual r values are also displayed in the plot.



*The following abbreviations have been used: crp = c-reactive protein, alb = albumin, hp = haptoglobin, tc = total cholesterol, tg = triglycerides, glu = glucose, fru = fructosamine, alt = alanine amino transferase, ast = aspartame amino transferase, gt = gamma glutamyl transferase, creat = creatinine, ca = calcium, alp = alkaline phosphatase, phos = phosphate, tibc = total iron binding capacity, fe-ptc = transferrin saturations, pot = potassium ldh = lactate dehydrogenase and urat = uric acid.

Figure 26. Hierarchical clustering - Dendrogram displaying results of hierarchical clustering of all 21 variables. Same abbreviations as in previous figures have been used. A cut-off of 580 in height was only selected as this clustered the data in three main groups which facilitates the visualization of the classes.

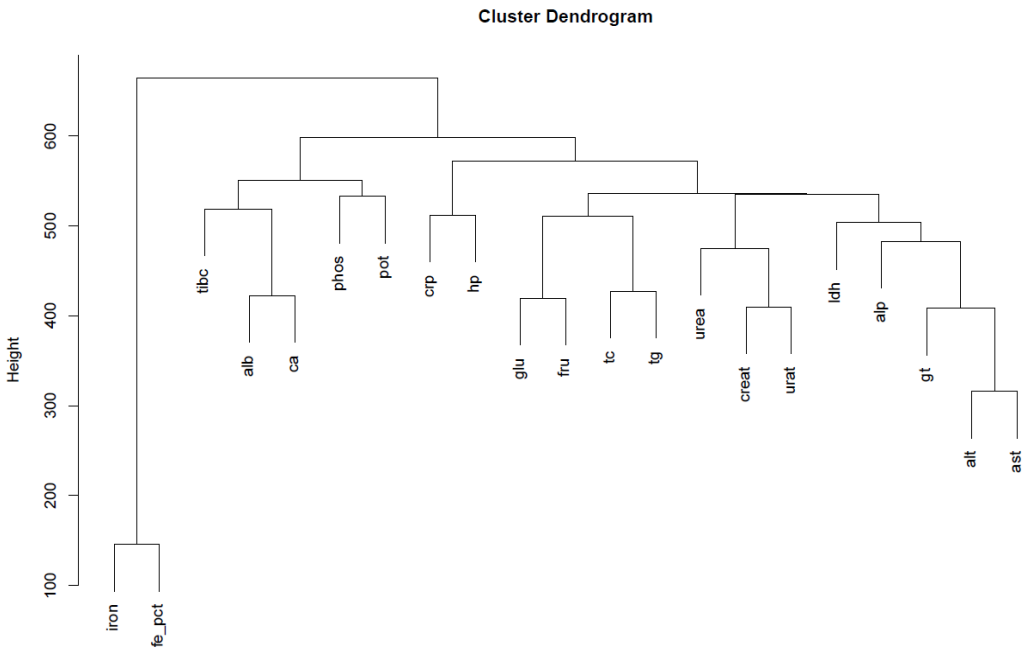
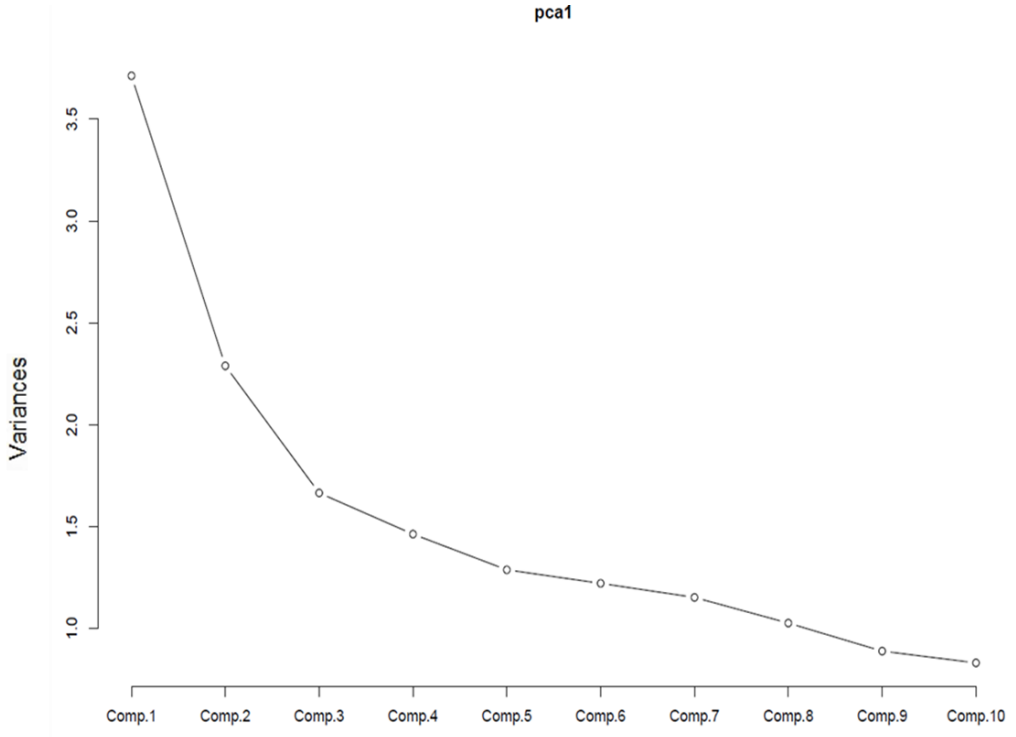


Figure 27. Principal component analysis Scree plot displaying variances of first 10 components versus the eigenvectors values with associated proportion of variance and cumulative proportions of the first 10 components with eigenvectors values > 1. The PCA analysis presented 12 components with eigenvector values > 1, however the R package only plots the 10 first components by default.



Proportion of variance %	18	11	8	7	6	6	5	5	4	4
Cumulative proportion %	18	29	37	43	50	55	61	66	70	74

Figure 28. PCA loadings for the first 12 components that presented eigenvector values > 1.

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12
CRP						-0.167	0.458	-0.422	0.647	0.212	-0.192	-0.217
ALT	0.308		0.443	0.189		0.148					0.244	
AST	0.303	-0.108	0.474	0.185		0.147			0.114		0.162	
GT	0.301		0.214		0.151	-0.152		-0.233	-0.395			-0.209
alp	0.271	0.113	0.118			-0.245	0.157	-0.186	-0.463		-0.377	-0.125
tg	0.319		-0.142		0.195	-0.123	-0.324		0.117	0.346	0.236	0.108
glu	0.260		-0.166	-0.144	0.431	0.292	0.306				0.198	
fru	0.270		-0.278		0.368	0.378	0.222					
alb		-0.123	-0.257	0.541				-0.240		-0.130	-0.124	0.157
hp	0.124	0.242		-0.228		-0.477	0.126			-0.131	0.231	0.415
tc	0.255		-0.174		0.186	-0.164	-0.353	0.323	0.247		-0.315	
pot	0.106		-0.137		-0.149	-0.335	0.334	0.506		-0.328	0.275	-0.352
creat	0.216		-0.222	-0.234	-0.464	0.199		-0.126			0.114	
urea	0.209		-0.222	-0.206	-0.404	0.205		0.133	-0.107	0.132	-0.115	-0.245
Ca	0.180		-0.283	0.414	-0.110	-0.246				-0.189	-0.166	0.304
iron		-0.595			0.138	-0.210					0.105	-0.241
phos		0.140		0.324	-0.110		0.230	0.248	-0.169	0.762	0.100	
tibc		0.274		0.357			-0.344	-0.183	0.124	-0.124	0.193	-0.566
ldh	0.249		0.253				-0.108	0.320	0.212		-0.491	
urat	0.339		-0.148	-0.146	-0.276		-0.203	-0.254			0.182	0.123
fe_pct		-0.640		-0.115		-0.161						

Table 5. Multivariate analysis of variance: mean values in each category of the given factors Sex, Education status, SES and Age. Same colour = no statistical difference between groups (tukey's range test). * Pr <0.0001, ** P<0.001, ***P>0.05.

	CRP	Alb	Hp	TC	TG	ALT	AST	GGT	Glu	Fru	Crea	Urea	Ca	ALP	Phos	Iron	TIBC	Fe	LDH	Uric	Pot
Sex	*	*	*	***	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
M	0.99	3.77	0.01	1.71	0.24	-0.84	-0.95	-0.87	1.61	0.75	4.46	1.65	0.87	0.97	0.01	2.87	4.07	-1.20	1.76	5.76	1.44
F	1.00	3.74	0.02	1.71	-0.02	-1.23	-1.14	-1.23	1.56	0.72	4.29	1.51	0.87	0.86	0.06	2.78	4.09	-1.31	1.75	5.49	1.43
Edu	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Low	1.02	3.74	0.07	1.75	0.21	-1.01	-1.02	-0.98	1.61	0.73	4.36	1.60	0.87	0.97	0.03	2.81	4.08	-1.27	1.78	5.64	1.44
Med	0.99	3.75	0.02	1.70	0.11	-1.03	-1.05	-1.05	1.58	0.73	4.37	1.57	0.87	0.91	0.04	2.83	4.08	-1.25	1.75	5.62	1.44
High	0.97	3.76	-0.03	1.68	0.02	-1.04	-1.05	-1.08	1.57	0.74	4.40	1.58	0.87	0.88	0.04	2.83	4.07	-1.24	1.73	5.63	0.08
SES	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Blue	1.00	3.75	0.03	1.70	0.12	-1.05	-1.05	-1.07	1.58	0.73	4.35	1.56	0.87	0.92	0.04	2.82	4.09	-1.27	1.76	5.60	1.44
White	0.98	3.76	0.00	1.71	0.11	-1.00	-1.03	-1.00	1.59	0.74	4.40	1.60	0.87	0.90	0.03	2.84	4.07	-1.23	1.74	5.66	1.44
U/M	1.05	3.74	0.05	1.71	0.14	-1.05	-1.02	-1.04	1.61	0.74	4.37	1.63	0.87	1.01	0.06	2.78	4.08	-1.30	1.80	5.66	1.43
Age	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
<40	0.98	3.78	-0.05	1.59	-0.04	-1.10	-1.09	-1.24	1.52	0.72	4.37	1.51	0.87	0.85	0.07	2.45	4.09	-1.24	1.70	5.59	1.42
40-50	0.96	3.75	0.02	1.71	0.11	-1.01	-1.06	-1.01	1.58	0.73	4.37	1.55	0.87	0.86	0.03	2.82	4.09	-1.26	1.73	5.61	1.44
50-65	1.00	3.74	0.06	1.78	0.22	-0.95	-1.00	-0.90	1.62	0.75	4.38	1.62	0.87	0.97	0.02	2.83	4.08	-1.25	1.79	5.67	1.44
>65	1.05	3.72	0.08	1.80	0.24	-1.04	-0.98	-0.91	1.65	0.76	4.40	1.70	0.87	1.04	0.02	2.80	4.06	-1.26	1.82	5.69	1.45

*The following abbreviations have been used: crp = c-reactive protein, alb = albumin, hp = haptoglobin, tc = total cholesterol, tg = triglycerides, glu = glucose, fru = fructosamine, alt = alanine amino transferase, ast = aspartame amino transferase, gt = gamma glutamyl transferase, creat = creatinine, ca = calcium, alp = alkaline phosphatase, phos = phosphate, tibc = total iron binding capacity, fe-ptc = transferrin saturations, pot = potassium ldh = lactate dehydrogenase and urat = uric acid.

d. Discussion

To explore the reciprocities between clinically meaningful blood biomarkers, I aimed to identify commonly measured biomarker groups into novel groups via their interplay with each other and the external environment. The statistical approach showed that these markers strongly correlate and cluster with other markers in similar and related body functions and metabolic pathways. Evaluation of the impact the external environment had on the internal environment indicated that lower level of education, older age and male sex were associated with higher biomarker levels that might indicated poorer health outcomes. There was a significant difference observed in the internal environment of individuals with different education levels, sex and age based on different strengths of correlations observed between the markers.

Internal environment

Serum biomarkers which are related to similar pathways and indicate adjacent body functions, such as the pairwise AST and ALT, Iron and FE-PCT, fructosamine and glucose and TC and TG, correlated and clustered strongly together. This was expected and in line with current understanding (250, 251). AST correlated with LDH, both enzymes are expressed in muscles and other body tissues and are released to blood stream where there is tissue damage. Specifically, the AST and LDH biomarkers, together with creatinine kinase and cardiac Troponin, are considered cardiac enzymes utilised for diagnosis of myocardial infarction (252-254). Albumin and Calcium markers showed a positive correlation and clustered together. The association observed between albumin and calcium may illustrate the interaction between those markers given that calcium in serum is bound to proteins, principally albumin. Total calcium measurement is usually corrected for by albumin (255). Finally, uric acid (Urate) correlated strongly with AST, ALT, GT, creatinine and triglycerides. Current literature suggests that elevated uric acid is an independent risk factor for non-alcoholic fatty liver disease and metabolic syndrome, even after adjustments for relevant confounders (256, 257).

For visualisation and interpretation purposes, hierarchical clustering separated the data into three main groups based on the selected height cut-off. One differentiated group contained only iron and its measure of saturation (Fe_pct), consistent with the strong correlations result for those markers. The second medium sized group contained TIBC, phosphate, potassium, albumin and calcium. The latter two were also found to correlate strongly and are related given that albumin carries calcium in blood. This association was further strengthened, as these two serum markers also produced their own cluster at the height of 490 on the dendrogram (Figure 26). Serum calcium levels are influenced by the parathyroid hormone, of which phosphate is also part, providing some explanation to the weak clustering observed between the two variables (258). Albumin has also been associated with TIBC as an indicator of inflammation in a malnutrition inflammation score (259). The largest cluster contained many of the commonly assessed metabolic and organ specific function markers (energy metabolic, inflammation, liver functions and kidney functions screening markers). Within this group, markers were most closely associated to similar metabolic pathways or body functions, hence indicating a complex interwoven internal environment, in which changes in one metabolic/organ pathway suggest consequence on another. Current literature also states that high circulating lipids may lead to insulin resistance, non-alcoholic fatty liver disease and poorer renal functions (260, 261). The results of the hierarchical clustering give a statistical basis to this understanding. The principal component analysis presented 12 components that accounted for 80 % of the variability of the data which did not indicate a variable or group of variables that accounted for the large variance in the dataset. The loadings of the components showed components enriched for most of the markers and components presenting few markers. Thus, this further strengthens the idea of an intertwined complex internal environment. Therefore, it is suggested that the classical epidemiological approach of study single markers has been useful to tackle specific biological question interrogating specific metabolism pathways, but it may not be sufficient to understand fully the different metabolic pathways interactions.

External environment

The results from the MANOVA analyses pointed out that males from lower education have higher serum biomarker levels that suggested associations with worse health outcomes. These findings are in line with large epidemiological and public health studies over the years (61, 62, 262-265). These studies suggest that healthy risk behaviours differ between levels of education, despite the public awareness of the implications about smoking, alcohol and exercise (263, 265). Therefore, individuals with lower education and lower SES are thought to be more likely to take part in risky health behaviours. Moreover, cohort studies presented worse health outcomes for individuals with lower SES, even after adjusting the analysis for health behaviours as confounders. This suggests that the interaction between SES and health is multifactorial, affect by life-style behaviours, education, environment (rural – urban) or access to exercise (264). Similar results were seen in the correlation analysis by education. In Figure 23, the correlation between glucose and fructosamine, a measure of glycaemic control, decreased as the level of education increased. Consequently, suggesting that other external, including behavioural factors, influence the fundamental internal environment of the body. This suggests the tight interaction between the different components of the exposome (external, general external and internal), even though the current study only investigated a small subset of the full exposome. When exploring the biomarkers by gender, ALP showed stronger correlations between some biomarkers in females, but not in males. ALP, as explained previously, is non- specific marker used as marker of liver function and bone turn over (197, 266). After menopause, oestrogens levels decrease whilst increasing women's cardiovascular risk , along with a decrease in bone mineral density increasing risk of osteoporosis (267, 268). This has been associated with raised levels of ALP (266). The results of the correlation analysis support this observation, a stratification of women pre-and post-menopausal could have helped to confirm this finding, however information on menopausal status was not available in AMORIS.

Future steps

Even though the principal component analysis did now show any remarkable results, future studies should aim to explore novel strategies to reveal clusters of markers that contain all the variability inherent to biological data by removing the noise and uninformative data, this in turn could allow better accuracy in the assessment of the impact of the external environment on the internal environment. However, it needs to be noted that there may be a strong interplay between the routinely collected markers shown in the current analyses. The internal biomarkers showed strong interconnections, beyond their established clinical utility, suggesting potential a reductionist approach in the current clinical use of the health biomarkers.

The findings of these exploratory analyses of a subset of the exposome confirmed tight synergy between the internal and external factors, which suggests the importance of exploring the complete exposome in the context of healthy ageing. Exposome studies should aim to assess exposures in different SES and education population strata and longitudinally along pertinent periods of human development.

Strengths and limitations

The major strength of this study lies with the large cohort and its external and internal validity. All participants were from the greater Stockholm area. All biomarker analyses were performed in the same laboratory. All individuals at the time were generally healthy and referred to the laboratory for a health check-up or were outpatients. However, any healthy cohort effect would not influence the internal validity of the study. Moreover, this study used a wide range of biomarkers and assessed them as individuals, not placing them in presumptuous groups. The AMORIS cohort is representative of the Swedish population (Table 1) and the subset of individuals studied in this project is comparable to the overall AMORIS population, however there is a slightly higher representation of the male population (Table 4). Even though there is a slightly greater proportion of employed subjects in

AMORIS, in comparison to the general population can be considered generalizable to another healthy western populations. During the study period, the all-cause mortality was about 14% lower in the AMORIS population than in the general population of Stockholm County when taking age, gender, and calendar year into account. Nevertheless, this healthy cohort effect would not affect the internal validity of our study and it is also likely to be minor since it has been shown that the AMORIS population is similar to the general working population of Stockholm County in terms of SES and ethnicity (167, 168). The main limitation of the study is the partial availability of information on life-style and environmental factors; the lack of these external exposures compromised the characterisation of the full exposome, however the results are relevant as an exploratory approach of a subset of the blood exposome.

Other limitation of this study is that the markers were measured at one point in time, a single measurement will not consider temporal variation. Exposome studies argue that if limited measurements are to be taken then they should be taken during sensitive periods of human development; such as during antenatal development, infant and puberty years (240).

Conclusion

I observed a complex synergy between the biomarkers and could not identify a single/group of markers that had a higher weighting over another that could account for the variance of the data. Lower education, older age and male sex were found to be markers associated with poorer health outcomes apart from ALP, which was more prominent in the female environment. The findings of this study indicate that using single markers to assess the internal environment may not represent the current underlying biological mechanisms, and may be susceptible to bias results.

Therefore, this exploratory study on a subset of the internal-external exposome (21 internal and 4 external markers) illustrates the importance of the further exploration of the exposome approach when measuring exposure impact and

health outcomes. Future studies should shift from single internal marker assessment to multiple markers at different time points, including multiple metabolites and external and specific external markers, when assessing an outcome such as cancer, which would capture a higher percentage of the heterogeneity studied allowing for a better characterisation of the disease. This would in theory contribute to better disease prediction of complex, multifactorial diseases – which in turn could lead to better public health strategies to prevent chronic diseases.

IV. Project B: Metabolic profiles in AMORIS

The findings of this project were presented as a poster at the NCRI Conference in 2015, where it was shortlisted for the BACR Hamilton Fairley poster prize (Appendix I). It was also presented as an oral presentation at the Big Data in Biology and Health conference in 2017 (Appendix II). The findings of this project are currently under review with British Journal of Cancer.

a. Rationale

The characterisation of the subset of the blood exposome in AMORIS in project A, exploring some internal established health markers, external demographics and socio-economics factors, confirmed a tight synergy between the internal and external factors, which supports the importance of using an exposome approach in the context of health.

Moreover, in 2013, Professor Christopher Wild proposed the assessment of the exposome in relation to cancer to elucidate putative risk factors associated with carcinogenesis for early disease detection (108). Blood metabolites, molecules capable of discovering causes of disease whilst channelling the internal exposome (117), have been widely explored as markers of exposure and susceptibility. For example, in cancer, increasingly seen as a metabolic disease, blood metabolites have contributed to identify groups in the population at greater risk of disease (118, 269-271). Furthermore, with the implementation of the "meet-in-the-middle approach", meaning that metabolites could be used as markers of exposure but also as markers of effect, the blood exposure has also been proven useful as a marker of effect or preclinical response to exposure (95). This was demonstrated in three pilot studies from the European Prospective Investigation into Cancer and Nutrition cohort (EPIC) on colon, breast and hepatocellular cancer outcome (244, 272). These studies established the role of blood exposome as a powerful tool to unravel cancer susceptibility in a population, which ultimately could lead to an effective stratified medicine.

However, is still uncertain how precisely metabolites can classify individuals according to cancer risk, given the heterogeneity, confounding factors and noise inherent to the biological data (120, 273). Therefore, novel statistical strategies, to explore multiple serum markers in relation to cancer and mortality, are needed to build efficient stratification models that reduce the phenotype space to its predictive components, which could potentially improve the diagnostic protocols in clinical practice (274). Recent studies have investigated different approaches for risk stratification For example, *Assi et al*. performed a partial least squares analysis to investigate metabolic profiles in relation to hepatocellular carcinoma (272). *Shann et al.* explored PCA stratification of individuals for CVD risk based on metabolic profiles (275), whilst *Lacey at al.* applied latent class analysis to cluster multiple external and general external exposures for pain scoring (276). In some of the studies, the association of the blood markers and the outcome of study was not strong enough to allow for disease stratification, however those markers could be still be used as markers of disease susceptibility.

Therefore, with the overall goal of investigate statistical methods to classify individuals based on their underlying risk of developing cancer and risk of increasing mortality, I conducted an exploratory data driven approach utilising routinely collected standard of care serum markers to study susceptibility to cancer and death in a well-defined cohort from the AMORIS study. More specifically, the study was designed to explore population heterogeneity and cancer susceptibility to investigate the capabilities of the LCA latent class serum metabolic profiles as possible risk stratification tools for cancer and mortality. .

b. Methods

i. Study population

This study utilises individuals from the Apolipoprotein Mortality-related Risk (AMORIS) database, as described above. For this project, I included all individuals aged 20 or above with baseline measurements of the following 19 standard of care metabolites (n=13,615): Glucose (mmol/L), Fructosamine (FAMN) (mmol/L), Total

Cholesterol (TC) (mmol/L), High Density Lipoprotein (HDL) (mmol/L), Low Density Lipoprotein (LDL) (mmol/L), Triglycerides (TG) (mmol/L), Apolipoprotein A-1 (ApoA-I) (g/L), Apolipoprotein B (ApoB) (g/L), Alanine Amino Transferase (ALT) (IU/L), Aspartate Amino Transferase (AST) (IU/L), Gamma-glutamyl transferase (GGT) (IU/L), Creatinine (µmol/L), Albumin (g/L), Leukocytes (WBC) (10^9 cells/L), C-reactive protein (CRP) (mg/L), Serum Iron (FE) (µmol/L), Total Iron Binding Capacity (TIBC) (mg/dL), Phosphate (mmol/L) and Calcium (mmol/L). All measurements were measured on the same day, using fully automated methods with automatic calibration performed on fresh blood samples. The blood metabolites included in the analysis were all the standard serum markers available from routine health check-ups, selected to represent main metabolic pathways in an exploratory exercise (Figure 20). The panel of biomarkers includes a high representation of markers of the energy metabolism with the aim of establishing the optimal markers to represent that specific pathway, given the abundance of these markers in the CALAB database.

Most of the markers comprised in the analysis have been previously studied individually in AMORIS, however no systemic integrative approach to examine the metabolic markers interactions and susceptibility to cancer has been conducted to date (Table 1) (169-184). All participants were free from cancer at time of study entry and none were diagnosed with cancer within the first three years of follow-up to avoid reverse causation.

The main outcomes were first cancer diagnosis, as registered in the National Cancer Register using ICD-9 for the years 1987-1992, ICD-O/2 for years 1993-2004 and for year 2005 onwards has been coded in ICD-O/3, and mortality. As secondary outcomes, I explored those cancer types for which there were more than 30 events during follow-up. Likewise, cancer mortality and overall mortality were explored. Follow-up time was assessed specifically for each of the outcomes studied. For cancer diagnosis, follow-up time was defined as time from blood drawn until date of first cancer diagnosis, death, emigration or study closing date (31st of December

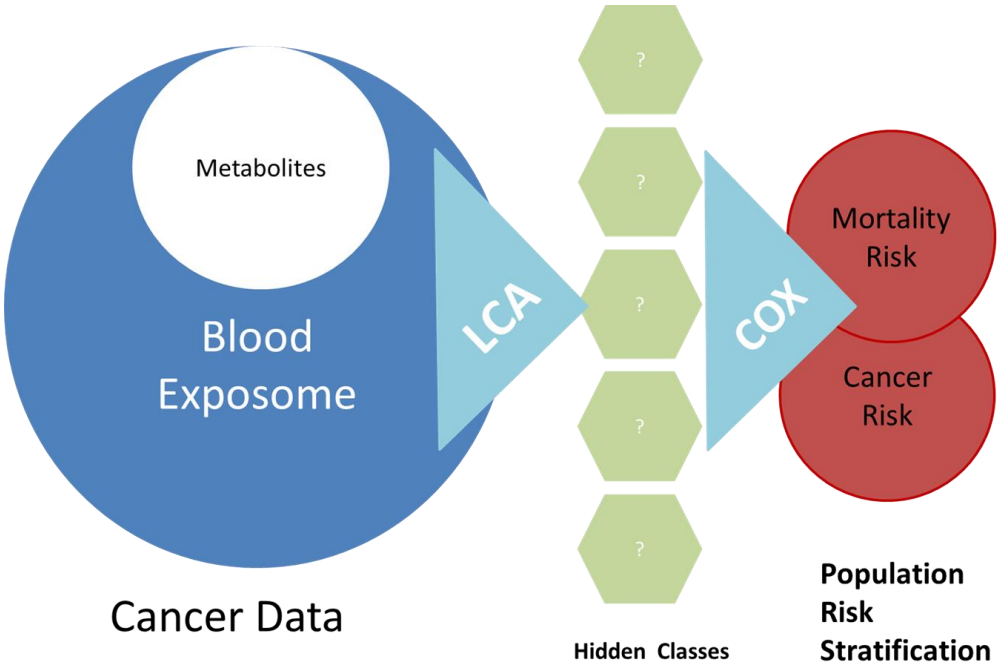
2012), whichever occurred first. The follow-up time for death was described as time from blood drawn until date of death, emigration or study closing date (31st of December 2012), whichever occurred first.

Information on the following potential confounders was also incorporated: age, sex, education status and comorbidities. The latter was quantified using the Charlson Comorbidity Index (CCI) calculated based on data from the National Patient Register. The CCI comprises 17 disease categories, all assigned a weight. The sum of an individual's weights was used to create the CCI ranging from no comorbidity to severe comorbidity (0, 1, 2, and ≥3) (277).

ii. Statistical Analysis

The statistical pipeline in this project is divided into three main analyses to explore the capabilities of multiple markers of the blood exposome, when reduced to informative metabolic profiles, to stratify a population by cancer risk and mortality risk. The analyses comprising the statistical pipeline were the following: (1) **descriptive statistics** of the study population, including data distribution and correlations, to explore the dataset and observe potential collinearity of the serum markers, (2) **latent class analysis** to characterise different classes of individuals based on their biomarker profiles, also evaluating intrinsic associations between those biomarkers and (3) **multivariable cox regression analysis** to examine whether the LCA classes, based on the panel of serum biomarkers, are associated with long-term risk of cancer and risk of all-cause and cancer-specific death (13, 128, 156). Figure 29 illustrates the methodological approach conducted in this project.

Figure 29. Methodological approach. Innovative avenue to explore cancer susceptibility in a well-defined cohort. This represents a shift from the classical targeted hypothesis driven approach to an exploratory data driven approach.



Furthermore, to perform the above-described **LCA analysis,** I determined the optimal number of LCA-derived classes by executing step-wise models with different numbers of classes, starting with the null model and allowing for one extra class in each model until reaching the total number of biomarkers in the data whilst the model kept converging into a local maximum of the likelihood function. The criterions used for model selection (Akaike information criterion (AIC), Bayesian information criterion (BIC) and Chi-squared (X^2)) were evaluated to estimate the best goodness of fit model and to define the optimal number of LCA-derived metabolic classes that characterised the dataset. Once the number of latent classes was established, I obtained the values of the parameters for the selected model. To identify which sets of biomarkers predominantly explained each latent class, how the classes were distributed across the study population and which individuals were allocated to each class, I assessed the conditional probabilities, mixed proportions and class memberships of the best fitted latent class model.

For the LCA categorical packages, *proc LCA* and *poLCA*, the number of latent classes was defined by the minimum value for BIC, AIC and X^2. For *MCLUST*, the LCA package for continuous data, BIC and AIC provided the best fitting model and best number of clusters for the dataset using maximum criteria (160, 162-164).

Once each subject was assigned to its LCA-derived metabolic class, I then conducted the final step of analytical pipeline, the **multivariate Cox proportional hazards regression analysis** to examine whether the LCA-derived metabolic classes were associated with long term risk of overall cancer as well as specific cancer types. In addition, I evaluated how the classes were associated with all cause-death and cancer-specific death. All models were adjusted for age, sex, and CCI. Education status was excluded of the analysis because of missing values. I performed a sensitivity analysis using age as a time-scale, as age is potentially a strong confounder in cancer. Moreover, Schoenfeld residuals were tested to ensure the proportional hazard assumption of the Cox regression analysis.

Finally, to assess the prediction capabilities of the LCA-derived metabolic classes for cancer, cancer death and overall death in comparison with standard health single biomarkers, c statistics were calculated for the LCA metabolic profiles compared to Total Cholesterol, Glucose and Gamma Glutamyl Transferase as single biomarkers.

The above explained statistical pipeline followed a multistage approach, which was developed using different analytical strategies. The analyses leading to this final statistical pipeline are described in more detail below.

First, the distribution of all biomarkers and demographic variables was evaluated using **descriptive statistics** including the mean, standard deviation and frequencies. Furthermore, histograms were plotted to identify the crude distribution of each biomarker. Pearson correlation coefficients were calculated to measure the strength of association between the biomarkers in the total study population as well as stratified by cancer outcome.

Second, **latent class analysis** was performed using different formats of the values of the biomarkers to establish the optimal format of the data to utilise LCA:

- Model A- Categorisation of the values, based on standard clinical cut-offs
- Model B- Z-transformation of quartiles values
- Model C- Z-transformation of continuous values

Model A and B were performed using the *proc LCA* package (162) in SAS and the *poLCA* package (163) in R. To allow for analysis of biomarkers as continuous variable, the *MCLUS*T package (164) in R was used. As a validation exercise, the study cohort was randomly split into a training (2/3) and test (1/3) datasets and LCA was executed for model A in both sets.

Finally, after selecting the best input format for the biomarkers in the LCA, a sensitivity analysis was conducted following the statistical pipeline, running **LCA followed by COX regression analysis to assess cancer risk** in time scale, by exploring different panels of serum markers:

- Model D- All markers including four established lipids ratios (23)
- Model E- All markers except the lipid related markers (13)
- Model F- All markers including TC and TG, but excluding the rest of lipids markers (15)
- Model G- All markers including the established ratios log(TG/HDL) and ratio ApoB/ApoA ratios, but excluding the rest of lipids markers (15)

After the sensitivity analysis, a final **Model H** was obtained which guaranteed the optimal performance of the statistical pipeline proposed for this particular project.

This exploratory multistep approach was employed to ensure an effective characterisation of individuals for cancer risk and mortality, whilst allowing for an easy biological/clinical interpretation of the results, given the novel implementation of the latent class analysis in this specific biomedical setting. The results section of

this project will cover all the different models studied. However, detailed interpretation is only provided for the final Model H.

Data management was performed using SAS version 9.4 (SAS Institute, Cary, NC, USA) and data analysis was conducted in SAS and R version 2.13.2 and R studio 3.3.2 (R Foundation for Statistical Computing, Wien, Austria) (248, 249).

c. Results

The tables and figures resulted from the analyses are displayed at the end of this section to facilitate the flow of information.

i. Characteristics of the study population

A total of 1,956 individuals (14.37%) developed cancer after at least 3 years of follow-up, including 655 breast and genito-urinary cancers, 330 cases of digestive cancer, 133 cases of respiratory cancers and 129 lymphatic and hematopoietic cancers, during a mean follow-up time of 16.6 years. 3,158 participants (23.20%) died during a mean follow-up of 17.3 years, comprising 706 cancer-specific deaths. About 50% of the individuals were in the 40-60 age strata, with high representation of participants older than 60 years of age, and there was a significant difference in age between individuals with and without cancer. The majority of participants were in the high SES status and middle to high educational status and presented no comorbidities at baseline. In general, the cancer subgroup showed slightly higher values for most of the biomarkers compared with the cancer-free group. For instance, high values of TC appeared in 32.5% of the cancer population versus 27.5% of cancer free population. Overall, 30% of the total population has high values for TC, TG and LDL biomarkers. Study population characteristics by cancer status are illustrated in Table 6. The descriptive statistics including clinical cut-offs and quartiles values for these biomarkers are listed in Table 7.

To evaluate the distribution of the data, histograms were created to identify the crude distribution of each of the markers. Most of the serum markers presented a

normal distribution of the data (Figures 30 – 32) and few markers showed a skewed distribution as illustrated in Figure 33. For some of the markers, a third of the population showed clinically abnormal values, as observed in table 6 (Figures 30 and 33). Pearson's correlation coefficients presented strong correlation between the different biomarkers in the lipid metabolism (TC, LDL and ApoB (r>0.7); HDL and ApoA-I (r>0.8)). The liver enzymes ALT and AST (r>0.8) also identified a strong correlation, as already observed in project A. The correlations by cancer status produced similar results. The properties of the total dataset remained in the training and test dataset.

ii. Assessment of the different formats of the biomarkers to perform LCA

I explored three different models to determine which input format of the biomarkers would produce a better performance of the LCA. In parallel, I investigated if the poLCA and proc LCA packages implemented for LCA in the statistical languages R and SAS showed the same behaviour when running LCA. Moreover, a small validation exercise to assess the LCA outcome when splitting the dataset was also conducted for the model A.

MODEL A

Latent class cluster analysis was executed using the **dichotomised values** of the biomarkers based on the clinical cut-offs (defined in table 6 and 7) in *proc LAC* for SAS and *poLCA* for R. The null model was run for 1 to 12 classes. The likelihood function did not converge into a local maximum for classes 9 to 12. Using R, the BIC and AIC, the goodness fit estimators, indicated a need for three or four classes (Figure 34a). After four classes both predictors stabilised, which indicated that adding more classes would not result in more information. Also in R, X^2 presented a minimum of three classes (Figure 35). Similar results were obtained for both the training and testing set (Figure 34c, Figure 34d). Using SAS, the BIC and AIC showed a slow decrease on their values without reaching a minimum for any of the classes; nevertheless, the largest decrease of the values occurred between two to four

classes (Figure 34b). Consequently, two, three and four latent classes were considered as possible models for the dataset. For all three models, the class allocation of the observations, the class conditional probability for each biomarker and the latent mixing proportions were identical when running R or SAS.

The results for each model are presented in Tables 8-10. Each table displays the number of latent class with the correspondence mixing proportion and the class conditional probabilities of belonging to each latent class for each of the abnormal clinical values of the biomarkers. The results of every model are explained below.

LCA for two latent classes separated the study population into two groups (Table 8) with the following latent class mixing proportion: class 2 represented 64% of the data and the other 36% belonged to class 1. Considering the biomarker distribution within the classes, the probability for abnormal values of TC, TG, LDL and ApoB was enriched in class 1, while class 2 clustered individuals with normal values for all the biomarkers.

LCA for three latent classes (Table 9) showed that class 3 contained 55% of the data and the other two classes held the rest of the population. Class 3 classified normal values for all the biomarkers, while the probability of abnormal values for TC, TG, LDL and ApoB all clustered in class 2, similarly to the previous model. Class 1 had a high probability of abnormal values of TG, HDL, glucose and creatinine.

The LCA four class model separated the study population into three groups (class 1, class 2 and class 3) with abnormal lipids measurements and one group (class 4) most likely resembling the normal population (Table 10). The training and validation dataset showed similar results, using both SAS and R.

MODEL B

To reduce possible confounding effects of units or dimensions and to ensure normal distribution for all the variables, **LCA** was performed characterising the continuous

biomarkers into **standardised quartiles** in the latter *poLCA* package in R. The null model was run for 1 to 16 classes for the standardised quartiles. BIC and AIC estimators were less informative in this analysis (Figure 36a). The predictors declined smoothly without reaching a minimum. The higher decrease was between two to three, then both predictors stabilized after four classes, consistent with the previous analyses using clinical cut-offs. X^2 did not reach a minimum, but just increased the values with the number of classes (Figure 36b). Therefore, I also examined the results for two and three latent classes using the standardised quartiles (Table 11, Table 12); in this case the tables explore only the conditional probabilities for the lipid profile and glucose biomarkers given their significance in previous analysis. Moreover, the remaining markers in the analysis did not show any trend within the classes.

LCA for two latent classes indicated that the mixing proportions of the latent classes were balanced (Table 11). One of the groups had abnormal high values for TC, TG, LDL and ApoB, while the other class presented low values for those lipids and high for HDL and ApoA. Moreover, LCA for three classes revealed a much more uniform distribution of the classes (Table 12). Hence, the information appeared diluted between the standardized quartiles. In this case, one of the clusters had high values of TC, TG and LDL and low HDL while other cluster showed lower values for TC, TG, LDL and high HDL and the last class had normal values.

MODEL C

Finally, to allow for analysis of biomarkers as **continuous variables**, I ran LCA on a **continuous** dataset and a **standardised continuous** dataset using the *MCLUST* package in R. This package does not require a null model to estimate the optimal number of classes. The package prints the BIC graph with all the possible models and number of classes. The class specific probabilities and latent mixing proportions for the best BIC value can therefore be retrieved. The analysis for the continuous values for all the biomarkers showed a model with one unique cluster (Figure 36b).

Moreover, the continuous standardised data showed a model with nine clusters (Figure 36a).

After performing three different analyses using Latent Class Cluster Analysis using 19 biomarkers in different data formats, the analyses were consistent irrespective of the software used to conduct LCA. The results of each of the model can be interpreted as follows:

MODEL A

LCA for clinical cut-offs for two classes showed class 2 representing the normal population and class 1 corresponding to individuals with dyslipidaemia, defined as a disorder of lipoprotein metabolism, including lipoprotein overproduction or deficiency (high serum TC, TG, LDL and low HDL) (278, 279). LCA for three classes clustered individuals with dyslipidaemia in class 2, class 1 individuals indicated presence of components of the metabolic syndrome (high serum TG and glucose) (280). Cluster 3 presented individuals with normal values (55%of the dataset).

The LCA with four classes was less informative, because it separated the study population into three groups (class1, class 2 and class 3) illustrating abnormal lipids and class 4 with normal values. Thus, using clinical cut-offs three classes was the optimal number of clusters in the population based on the BIC, AIC and X^2 goodness fit estimators. This outcome was internally validated using the training and testing set.

MODEL B

Using LCA for the standardised quartiles LCA models**,** the three classes model presented one class with dyslipidaemia, another one with low lipid profile and a third one as the normal population. The 2-class model divided the population in those with abnormal values for TC, TG and LDL and those with a low lipid profile. Latent Class analysis with standardised quartiles thus resulted in a dilution of the information between the quartiles; showing a balanced distribution of the classes

within the study population and a polarised characterisation of individuals into low and high lipid profiles.

MODEL C

LCA with continuous and standardised continuous data was difficult to interpret and it was not possible to extract any information from the population when used. Continuous data analysis showed all the markers in one unique cluster, which did not reduce the dimension of the data or provided information about heterogeneity in the population. Moreover, standardised continuous data analysis presented 9 classes as the best model for the data, which similarly did not reduce the dimension of the data sufficiently.

The above results thus indicated that the LCA using clinical cut-offs characterised the population into biologically meaningful groups, while the standardised quartiles seem to stratify the population between high and low values for biomarkers. This may be because only a small proportion of the population had abnormal values for the biomarkers used according to the clinical cut-offs, whereas there was a more homogenous distribution of the population when using quartiles of the biomarkers (Figure 31, Figure 32). Also from a clinical point of view, keeping cancer risk as an outcome in mind, it seems reasonable to use the medical cut-offs. Furthermore, it has been suggested previously that scaling and normalising is not required for this method in comparison to other data reduction methods (281). Moreover, LCA was developed originally for categorical data analysis, and the implementation for continuous data was performed using a different R package *mclust* which might explain why the results for continuous data were very different to the results of clinical cut-off and quartiles. For both quartiles and clinical cut-off analyses, the models with number of classes estimated best for the data were similar, which confirms that this method is not affected by the scaling of the data. The difference between the results of the clinical cut-offs and quartiles analyses lies in the description or content of the different classes, which was based on the method used to create the categories. The utility of the data reduction methods is strongly driven

by the interpretability of the results in a given context, so there will always be a subjective component in the selection of the best model for the data, which will be based on previous knowledge and expertise in the area of research. Therefore, given the clinical and biological context of the study, clinical cut-offs presented the results that were most informative for this particular area of research.

This assessment confirms that clinical cut-offs values seem the best input format to perform LCA in a biomedical context. The results from Model A are clinically and biologically meaningful. Moreover, the goodness fit indicators of the model were more precise in model A, especially for the model fit indicator X^2.

iii. Sensitivity analysis of the panel of biomarkers for LCA and COX

After identifying that the clinical cut-offs were the best input format of the biomarkers for LCA, a sensitivity analysis was conducted by exploring different panels of serum markers.

The following analyses were conducted in the statistical analysis program R, given its faster performance in comparison with SAS and the consistent results for LCA in both platforms. As mentioned previously, the biomarker panel contained a high proportion of lipid markers (TC, HDL, LDL, TG, ApoA-1 and ApoB) with the intention of establishing those markers most representative of that pathway. In standard medical practice, TC and TG biomarkers are commonly measured in blood samples to assess the lipid metabolism, however recent publications present the ratios ApoB/ApoA-I and log (TG/HDL) as superior markers characterising the internal and external axis of the lipid component (genomic/environment) (187, 194, 282-284). Moreover, LCA requires the principle of local independence between the covariates analysed (159). The correlation analysis showed collinearity for these markers, which might explain the strong lipid burden seen in the previous results. Therefore, the panel of biomarkers was further investigated to include only the best possible markers to stratify the population for cancer risk, specifically concentrating on lipids, by running LCA followed by a multivariate COX regression analysis adjusted for age,

sex, education status and CCI using calendar-time as a time scale. Four different models were investigated: MODEL D which includes the full panel of biomarkers adding four established lipid ratios: ApoB/ApoA-I, Total cholesterol/HDL, LDL/HDL and Log (triglycerides/HDL) (23), MODEL E containing all markers excluding the lipid factors (13), MODEL F based on all markers but only TC and TG of the lipid pathway (15) and finally MODEL G which contained all markers including the log(TG/HDL) and ApoB/ApoA-1 ratios, but excluded the rest of lipids markers (15). Given the extensive analytical pipeline, the description of the results is summarised below.

MODEL D

LCA was executed using dichotomised values of the full biomarker panel and four lipid markers (ratios (ApoB/ApoA-I, Total cholesterol/HDL, LDL/HDL and Log (triglycerides/HDL))) based on clinical cut-offs in poLCA in R. The null model was run for 1 to 12 classes. The likelihood function did not converge into a local maximum for classes 11 and 12. The BIC and AIC, the goodness fit estimators, indicated a need for three or four classes and X^2 presented a minimum of four classes (Figure 38a). Consequently, four classes were considered the best model for the dataset. Table 13 presents the results for four classes. The LCA separated the data into four groups: the largest represented the normal values for the markers (52%), class 3 was enriched with TC, LDL, ratio ApoB/apoA-1 (20%), class 2 was largely enriched in the lipid markers, glucose (16%) and class 1 (12%) with abnormal values for liver enzymes, ALT and elevated GGT. Once each subject was assigned to its LCA-derived metabolic class, a multivariate COX regression analysis to assess whether the metabolic profiles were associated with long term cancer risk was performed. Class 1 represented by the altered liver enzymes showed a significance increase of over all cancer risk (HR: 1.19 (95%CI: 1.02 - 1.37)) for the individuals allocated in that class. The COX regression analysis results for the four class-model are listed in Table 14.

MODEL E

To explore the biomarker panel whilst excluding the extensive lipid component, LCA was performed on the remaining 13 markers. The null model was run for one to 12 classes. The likelihood function did not converge into a local maximum for classes 9 to 12. The BIC and AIC, the goodness fit estimators, did not presented a clear minimum and X^2 presented a total minimum of three classes (Figure 38b). Consequently, three classes were considered the best model for the dataset. Table 15 presents the results for three classes. The LCA separated the data in three groups: the largest represented the normal values for the markers (78%), class 3 showed high Glucose and high liver enzymes ALT, AST and GGT (15%) and class 1 (7%) clustered individuals with abnormal values for iron markers. The multivariate regression analysis did not present any statistical significant association with long term risk of cancer for any of the classes in comparison with the normal class 2 (Table 14).

MODEL F

Following the exploratory analysis of the non-lipid related markers, LCA was executed for all the markers including only TC and TG from the lipid component (15). The null model was run for one to 12 classes. The likelihood function did not converge into a local maximum for classes 8 to 12. The BIC and AIC, the goodness fit estimators, did not presented a clear minimum and X^2 presented a minimum of three classes (Figure 38c). Consequently, three classes were considered the best model for the dataset (Table 16). Class 1 included individual with normal values for the markers (68%), while class 2 clustered 25% of the population presented with abnormal values for TC, TG and ALT, GGT from the liver metabolism. Class 3 showed low levels for iron and TIBC (7%). Class 2, represented by lipids and the altered liver enzymes, showed a significant increase of over all cancer risk (HR: 1.16 (95%CI: 1.04 - 1.28)) for the individuals allocated in that class (Table 14).

MODEL G

LCA was performed on all the markers including only the ratio log (TG/HDL) and ratio ApoB/ApoA-1 for the lipid pathway (15). The null model was run for one to 12 classes. The likelihood function did not converge into a local maximum for classes 8 to 12. The BIC and AIC, the goodness fit estimators, indicated a need for three or four classes and X^2 presented a minimum of four classes (Figure 38d). Consequently, four classes were considered the best model for the dataset. LCA clustered individuals with normal values for all the markers in class 1, representing the largest population (63%). Participants allocated in class 2 manifested high values for the lipid ratios log (TG/HDL) and ApoB/ApoA-1 being 23% of the population. Class 3 classified individuals with high values for the enzymes ALT, AST and GGT (9%), and class 4 clustered individuals with low values for iron and TIBC (6%) (Table 17). Multivariate cox regression analysis showed that individuals with high abnormal values for liver enzyme class 3 had a higher risk of overall cancer in comparison with the normal class 1 (HR: 1.29 (95%CI: 1.10 - 1.52)) (Table 14).

After conducting LCA followed by Multivariate Cox regression analysis to assess cancer risk in the four models that included diverse panels of biomarkers, the results showed that metabolic profiles classification based on MODEL D was mainly driven by the lipid biomarkers, which was expected given the large number of biomarkers from that metabolic pathway included in the analysis. Moreover, MODEL E, which represented the rest of the markers after removing the lipid panel, indicated the importance of the enzymes ALT, AST and GGT, common markers of liver disease, glucose and iron metabolites when classifying individuals to assess hidden heterogeneity based on blood metabolites. Furthermore, more interesting scenarios including all the markers and a small subset of the lipids (TC and TG for MODEL F and ratio log (TG/HDL) and ratio ApoB/ApoA-1 for MODEL G) indicated that lipid markers, liver enzymes and iron metabolites carried the main variability within the population. MODEL F allocated the individuals in three independent classes, one clustering abnormal values for TC, TG, ALT, AST and GGT. MODEL G

more precisely allocated four subgroups of individuals with diverse metabolic profiles each driven by markers related to similar pathways.

For all the models examined, the X^2 goodness fit indicator reached a minimum facilitating the estimation of the number of classes inherent to each model. Multivariate COX regression analysis indicated that the individuals classified within the class with abnormal values for the enzymes ALT, AST and GGT presented higher risk of overall cancer in comparison with the normal class for three of the models studied (Table 14). Figure 39 summarizes the sensitivity analyses results for the four models investigated.

Therefore, the sensitivity analyses confirmed that the panel of markers included in Model F and Model G were more appropriate to study the population heterogeneity hidden in blood metabolites in the AMORIS subcohort, given that markers representing the main metabolites were included whilst avoiding an excessive burden of the lipids on the analysis. LCA for MODEL G presented a clearer cluster of the individuals and the ratios seemed to represent more accurately the internal and external component of the lipid pathway (187, 194, 285).
This model was also supported by a small subanalysis where I characterised the study population based on the definition of abnormal clinical values for each of the two lipid panels:

- For definition of clinically abnormal TC & TG: The study population was constituted by individuals with high lipids n=5704 (42%) and individuals with normal lipids n=7911 (58%)
- For definition of clinically abnormal ratio log (TG/HDL) and ratio ApoB/ApoA-1: The study population was defined by individuals with high lipids n=4723 (35%) and individuals with normal lipids n=8892 (65%)

The AMORIS cohort is characterised by individuals with high lipids representing 32% of the population with lipid measurements available and individuals with normal lipids being 68% of the cohort of participants with those markers available.

Thus, this exploratory multistep approach helped identify an effective risk stratification of individuals for cancer whilst allowing for an easy biological/clinical interpretation of the results. The following biomarkers, when dichotomised based on clinical cut-offs, were found to be the most effective input for the statistical pipeline performed in this project: Glucose (mmol/L), Fructosamine (FAMN) (mmol/L), ratio log(TG/HDL) and ratio ApoB/ApoA-1, Alanine Amino Transferase (ALT) (IU/L), Aspartate Amino Transferase (AST) (IU/L), Gamma-glutamyl transferase (GGT) (IU/L), Creatinine (µmol/L), Albumin (g/L), Leukocytes (WBC) (109 cells/L), C-reactive protein (CRP) (mg/L), Serum Iron (FE) (µmol/L), Total Iron Binding Capacity (TIBC) (mg/dL), Phosphate (mmol/L) and Calcium (mmol/L). This definitive scenario, MODEL H guarantees the optimal performance of the statistical pipeline proposed in this particular setting.

iv. Definitive input model for the statistical pipeline: MODEL H

Latent Class Analysis characterizes the study population into four metabolic profiles based on the above-mentioned Model H (Figure 40A, Figure 40B) (43). The class allocation of the observations (individuals), the class conditional probability of each biomarker and the latent mixing proportions were obtained when running poLCA package in R statistical language.

Table 18 and Figure 41 outline the LCA-derived classes with the estimated class population proportions, the class conditional probabilities of belonging to each latent class for each of the biomarkers and the biological interpretation of the LCA-derived classes. Figure 41 was included to allow easier visualisation of the LCA-derived classes. The four mutually exclusive classes characterised the population into metabolic profiles based on class conditional probabilities: (1) those with probabilities for all abnormal values of the markers under 0.3; therefore, considered the normal class (63% of population); (2) those with abnormal values for lipid markers (22%); (3) those with abnormal values for liver function markers (9%); (4) those with abnormal values for iron and inflammation metabolism (6%). Figure 40

illustrates each of the different profiles based on all the abnormal values for the markers.

A validation of the characterisation of the population performed with the Latent class methodology is outlined in Table 19. The baseline clinical characteristics of the individuals by LCA-derived metabolic classes (Table 19) replicate the results displayed in Table 18 for the class conditional probabilities.

LCA derived metabolic profiles in relation with cancer and mortality

I then investigated the capabilities of the four LCA-derived metabolic profiles as cancer susceptibility metabolic profiles in relation to overall cancer risk, specific cancer types risk, cancer mortality and overall mortality, assigning the reference level to the healthy metabolic profile Class 1 (Tables 20 - 21).

All metabolic profiles increased risk of cancer and mortality compared to Class 1. For instance, individuals in Class 3 (abnormal liver function profile) had a higher risk of overall cancer (HR: 1.28 (95%CI: 1.10- 1.50)), but also a worse cancer-specific survival and overall survival as compared to those in Class 1 (Tables 20 – 21). Class 2 (abnormal lipid profile) and Class 4 (abnormal iron markers and inflammatory) were positively associated with overall death, while Class 2 was also associated with cancer–specific death. The results were consistent for both time-scales (Tables 20 – 21).

When assessing the risk of specific cancer types, several patterns occurred (Table 20). Individuals in Class 2 (abnormal lipid markers) presented a higher risk of lymphatic and hematopoietic tissue cancer (HR: 1.72 (95%CI: 1.15 - 2.56)). There was a greater risk of digestive cancers in individuals in Class 3 (abnormal values of liver enzymes) (HR: 2.12 (95%CI: 1.54 - 2.91)), while individuals in Class 4 (abnormal iron markers and inflammation) were exposed to a higher risk of buccal and oral system cancers in comparison with the individuals in Class 1 (HR: 3.94 (95%CI: 1.38 - 11.30)) (Table 20).

Moreover, the connective tissue and endocrine glands cancer risk was higher in individuals grouped in liver metabolic profile (HR: 2.65 (95%CI: 1.00 - 7.02) and in participants belonging to the iron markers and inflammation (HR: 3.00 (95%CI: 1.11 - 8.11)). Similar associations were observed when using the age scale for the Cox regression model (Tables 20 – 21).

All the Cox analyses were adjusted for age, sex and comorbidities (CCI) using calendar-time as a time scale. Models with age as a time scale were adjusted for sex and CCI. Education status was excluded in these analyses because of the missing values. The test for Schoenfeld residuals showed that the hazard ratios were proportional for all the outcomes studied.

C statistics test were conducted for LCA-derived metabolic classes as well as Total Cholesterol, Glucose and Gamma Glutamyl Transferase as single biomarkers. These models were performed for cancer, cancer death and overall death as an outcome using calendar-time as a time scale whilst adjusting for age, sex and CCI. LCA metabolic profiles presented very similar, but slightly better prediction capabilities than standard health single biomarkers for the three outcomes, especially for cancer mortality and mortality (Table 22).

Table 6. Characteristics of the study population by cancer status. All the serum markers are dichotomized using standard clinical cut-offs. * Clinically abnormal cut-off values are highlighted for each biomarker.

	Total	**No Cancer**	**Cancer**
	N=13,615 (100%)	**N=11,659 (85.63%)**	**N=1,956 (14.37%)**
Age (years)			
Mean (SD)	51.91 (14.80)	50.86 (15.00)	58.14 (11.75)
Under 40	2951 (21.67)	2841 (24.37)	110 (5.62)
40-50	3550 (26.07)	3148 (27.00)	402 (20.55)
50-60	3065 (22.51)	2491 (21.37)	574 (29.35)
Above 60	4049 (29.74)	3179 (27.27)	870 (44.48)
Sex			
Female	7588 (55.73)	6636 (56.92)	952 (48.67)
Male	6027 (44.27)	5023 (43.08)	1004 (51.33)
Socio-economics Status			
High	6493 (47.69)	5416 (46.45)	1077 (55.06)
Low	5007 (36.78)	4368 (37.46)	639 (32.67)
Not employed or missing	2115 (15.53)	1875 (16.08)	240 (12.27)
Educational Status			
High	4313 (33.42)	3688 (33.40)	625 (33.57)
Middle	5495 (42.58)	4725 (42.79)	770 (41.35)
Low	3097 (24.00)	2630 (23.82)	467 (25.08)
Missing [a]	710 (5.21)	616 (5.28)	94 (4.80)
CCI			
0	12258 (90.03)	10520 (90.23)	1738 (88.85)
1	963 (7.07)	807 (6.92)	156 (7.98)
2	221 (1.62)	188 (1.61)	33 (1.69)
3+	173 (1.27)	144 (1.24)	29 (1.48)
Glucose (mmol/L)			
Mean(SD)	5.22 (1.53)	5.21 (1.53)	5.30 (1.53)
< 6.11	12223 (89.78)	10488 (89.96)	1735 (88.70)
≥ **6.11**	1392 (10.22)	1171 (10.04)	221 (11.30)
Fructosamine (mmol/L)			
Mean(SD)	2.09 (0.27)	2.08 (0.27)	2.10 (0.25)
< 2.6	13184 (96.83)	11291 (96.84)	1893 (96.78)
≥ **2.6**	431 (3.17)	368 (3.16)	63 (3.22)
Total Cholesterol (mmol/L)			
Mean(SD)	5.82 (1.17)	5.79 (1.18)	6.00 (1.13)
< 6.50	9774 (71.79)	8453 (72.50)	1321 (67.54)
≥ 6.50	3841 (28.21)	3206 (27.50)	635 (32.46)
HDL Cholesterol (mmol/L)			
Mean(SD)	1.54 (0.43)	1.54 (0.43)	1.52 (0.43)

< 1.03	1457 (10.70)	1231 (10.56)	226 (11.55)
≥ 1.03	12158 (89.30)	10428 (89.44)	1730 (88.45)
LDL Cholesterol (mmol/L)			
Mean(SD)	3.64 (1.06)	3.61 (1.06)	3.82 (1.04)
< 4.10	9345 (68.64)	8128 (69.71)	1217 (62.22)
≥ **4.10**	4270 (31.36)	3531 (30.29)	739 (37.78)
Triglycerides (mmol/L)			
Mean(SD)	1.44 (1.00)	1.43 (1.00)	1.48 (0.93)
< 1.71	10128 (74.39)	8716 (74.76)	1412 (72.19)
≥ 1.71	3487 (25.61)	2943 (25.24)	544 (27.81)
Apolipoprotein A-1 (g/L)			
Mean(SD)	1.44 (0.23)	1.44 (0.23)	1.43 (0.23)
< 1.05	328 (2.41)	278 (2.38)	50 (2.56)
≥ 1.05	13287 (97.59)	11381 (97.62)	1906 (97.44)
Apolipoprotein B (g/L)			
Mean(SD)	1.22 (0.35)	1.22 (0.35)	1.29 (0.34)
< 1.50	10902 (80.07)	9431 (80.89)	1471 (75.20)
≥ **1.50**	2713 (19.93)	2228 (19.11)	485 (24.80)
ALT (IU/L)			
Mean(SD)	29.02 (34.35)	28.95 (35.73)	29.41 (24.54)
< 50	12296 (90.31)	10546 (90.45)	1750 (89.47)
≥ **50**	1319 (9.69)	1113 (9.55)	206 (10.53)
AST (IU/L)			
Mean(SD)	22.84 (19.23)	22.70 (19.60)	23.64 (16.88)
< 45	13155 (96.62)	11271 (96.67)	1884 (96.32)
≥ 45	460 (3.38)	388 (3.33)	72 (3.68)
GGT (IU/L) *			
Mean(SD)	33.21 (48.12)	32.74 (48.09)	36.03 (48.21)
Normal (<18)	5511 (40.48)	4827 (41.40)	684 (34.97)
Normal high (18-36)	4983 (36.60)	4236 (36.33)	747 (38.19)
Elevated (36-72)	2098 (15.41)	1750 (15.01)	348 (17.79)
Highly elevated (>72)	1023 (7.51)	846 (7.26)	177 (9.05)
Creatinine (µmol/L) *			
Mean(SD)	79.65 (16.16)	79.38 (16.37)	81.26 (14.74)
Low	40 (0.29)	31 (0.27)	9 (0.46)
Normal	12088 (88.78)	10392 (89.13)	1696 (86.71)
High	1487 (10.92)	1236 (10.60)	251 (12.83)
Albumin (g/L)			
Mean(SD)	43.05 (2.82)	43.13 (2.83)	42.58 (2.72)
<35	28 (0.21)	23 (0.20)	5 (0.26)
>35	13587 (99.79)	11636 (99.80)	1951 (99.74)
Leukocytes (10^9 cells/L)			
Mean(SD)	6.52 (1.97)	6.49 (1.96)	6.65 (2.01)

<10	12956 (95.16)	11106 (95.26)	1850 (94.58)
≥ **10**	659 (4.84)	553 (4.74)	106 (5.42)
C-Reactive Protein (mg/L)			
Mean(SD)	5.86 (15.14)	5.82 (14.25)	6.16 (19.58)
<10	11858 (87.1)	10193 (87.43)	1665 (85.12)
10- 15	1196 (8.78)	993 (8.52)	203 (10.38)
15-25	265 (1.95)	223 (1.91)	42 (2.15)
25-50	200 (1.47)	167 (1.43)	33 (1.69)
>50	96 (0.71)	223 (0.71)	13 (0.66)
Iron (µmol/L) *			
Mean(SD)	18.13 (5.80)	18.13 (5.83)	18.11 (5.59)
Low	636 (4.67)	540 (4.63)	96 (4.91)
Normal	12512 (91.90)	10715 (91.90)	1797 (91.87)
High	467 (3.43)	404 (3.47)	63 (3.22)
TIBC (mg/dL) *			
Mean(SD)	0.39 (0.11)	0.31 (0.11)	0.31 (0.10)
Low	4067 (29.87)	3494 (29.97)	573 (29.29)
Normal	6650 (48.84)	5683 (48.74)	967 (49.44)
High	2898 (21.29)	2482 (21.29)	416 (21.27)
Phosphate (mmol/L) *			
Mean(SD)	1.07 (0.17)	1.07 (0.17)	1.05 (0.17)
Low	95 (0.70)	76 (0.65)	19 (0.97)
Normal	12796 (93.98)	10948 (93.90)	1848 (94.48)
High	724 (5.32)	635 (5.45)	89 (4.55)
Calcium (mmol/L) *			
Mean(SD)	2.38 (0.09)	2.38 (0.09)	2.38 (0.10)
Low	191 (1.40)	167 (1.43)	24 (1.23)
Normal	13195 (96.92)	11300 (96.92)	1895 (96.88)
High	229 (1.68)	192 (1.65)	37 (1.89)
Log (triglycerides/HDL) b			
mean(SD)	(-)0.19 (0.81)	(-)0.20 (0.82)	(-)0.14 (0.80)
< 0.5	11197 (82.24)	9618 (82.49)	1579 (80.73)
≥ 0.5	2418 (17.76)	2041 (17.51)	377 (19.27)
ApoB/ApoA-I b			
mean(SD)	0.87 (0.29)	0.87 (0.29)	0.92 (0.30)
< 1.00	9584 (70.39)	8347 (71.59)	1237 (63.24)
≥ 1.00	4031 (29.61)	3312 (28.41)	719 (36.76)
Life Status			
Alive	10457 (76.80)	9385 (80.50)	1072 (54.81)
Death	3158 (23.20)	2274 (19.50)	884 (45.19)
Cancer	1956 (14.90)	11659 (0.00)	1956 (100.00)
Follow up time (years)			
Cancer Mean(SD)	16.57 (4.74)	17.40 (4.16)	11.60 (4.93)

Death Mean(SD)	17.26 (4.29)	17.40 (4.16)	16.39 (4.89)

The following abbreviations have been used in Table 5: High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT) and Total iron binding capacity (TIBC).
a The missing values are not included in the percentage of the Educational Status categories
b Ratios are dimensionless
*Clinical cut-offs
The following cut-offs criteria was applied:
GGT reference interval:
Low [GGT < 36 IU/L]
Normal [36 IU/L ≥ GGT <72 IU/L]
High [GGT ≥ 72 IU/L]
Creatinine reference interval:
Men [Low ≤ 60, Normal = 60-100, High ≥ 100]
Women [Low ≤ 45, Norma l= 45-90, High ≥ 90]
Iron reference interval:
Men [Low ≤ 11, Normal = 11-31, High ≥ 31]
Women [Low ≤ 9, Normal = 9-30, High≥ 30]
TIBC reference interval:
Men [Low ≤ 0.257, Normal = 0.257-0.379, High ≥ 0.379]
Women [Low ≤ 0.246, Normal = 0.246- 0.391, High ≥ 0.391]
Phosphate reference interval:
Men [Low ≤ 0.7, Normal = 0.7-1.4, High ≥ 1.4]
Women [Low ≤ 0.8, Normal = 0.8-1.4, High ≥1.4]
Calcium reference interval per gender by age:
Men
[Age < 40, Low ≤ 2.22, Normal = 2.22-2.60, High ≥2.60]
[Age 40-60, Low ≤ 2.20, Normal = 2.20 -2.59, High ≥2.59]
[Age > 60, Low ≤ 2.19, Normal= 2.19 -2.58, High ≥ 2.58]
Women
[Age < 40, Low ≤ 2.17, Normal = 2.17-2.56, High ≥2.56]
[Age 40-60, Low ≤2.19, Normal = 2.19-2.60, High ≥2.60]
[Age > 60, Low ≤ 2.21, Normal = 2.21-2.60, High ≥2.60]

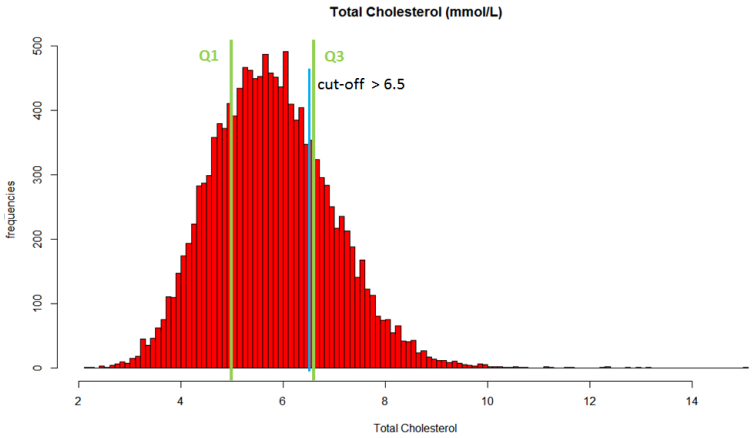
Table 7. Descriptive statistics of the data including clinical cut-offs and quartiles.

Biomarkers	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Cut-off low	Cut-off high	Units
Glucose	0.30	4.50	5.00	5.22	5.40	39.90	-	6.10	mmol/L
Fructosamine	1.22	1.94	2.06	2.09	2.19	6.10	-	2.60	mmol/L
Total Cholesterol	2.10	5.00	5.80	5.82	6.60	13.20	-	6.50	mmol/L
HDL	0.02	1.26	1.53	1.54	1.80	6.35	1.30	-	mmol/L
LDL	0.58	2.88	3.58	3.64	4.31	12.50	-	4.10	mmol/L
Triglycerides	0.20	0.80	1.20	1.44	1.80	17.10	-	1.71	mmol/L
ApoA	0.42	1.29	1.42	1.44	1.57	2.97	1.05	-	mmol/L
ApoB	0.23	0.96	1.19	1.22	1.43	5.05	-	1.50	mmol/L
ALT	2.40	16.80	22.80	29.02	32.99	2616.08	-	50.00	IU/L
AST	3.60	16.20	20.40	22.84	25.19	1238.15	-	45.00	IU/L
GGT	3.00	14.40	21.00	33.21	34.19	1546.49	-	36 -72	IU/L
Creatinine	34.00	70.00	78.00	79.65	88.00	656.00	45.00 - 60.00	90.00 - 100.00	µmol/L
Albumin	25.00	41.00	43.00	43.05	45.00	54.00	35.00	-	g/L
WBC	1.40	5.20	6.20	6.51	7.50	42.50	-	10.00	10^9cells/L
CRP	1.00	3.00	4.00	5.87	6.00	720.00	-	10.00	mg/L
Iron	3.00	14.00	18.00	18.13	21.00	62.00	9.00 - 11.00	30.00 -31.00	µmol/L
TIBC	0.04	0.24	0.30	0.31	0.37	0.98	0.246 - 0.257	0.379-0.391	mg/dL
Phosphate	0.40	1.00	1.10	1.07	1.20	3.20	0.70 - 0.80	1.40	mmol/L
Calcium	2.00	2.31	2.37	2.38	2.44	3.44	2.17- 2.22	2.56-2.60	mmol/L

*The following abbreviations have been used in Table 6: High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Figure 30. Histograms of the data distribution for TC. TC presents a normal distribution of the data. A) Blue line is the cut-off value, while green lines are the quartiles values. Approximate 30 % of the population has higher values for TC (cut-off > 6.5 mmol/L). B) Histograms by sex status (1=male, 2=female).

A)



B)

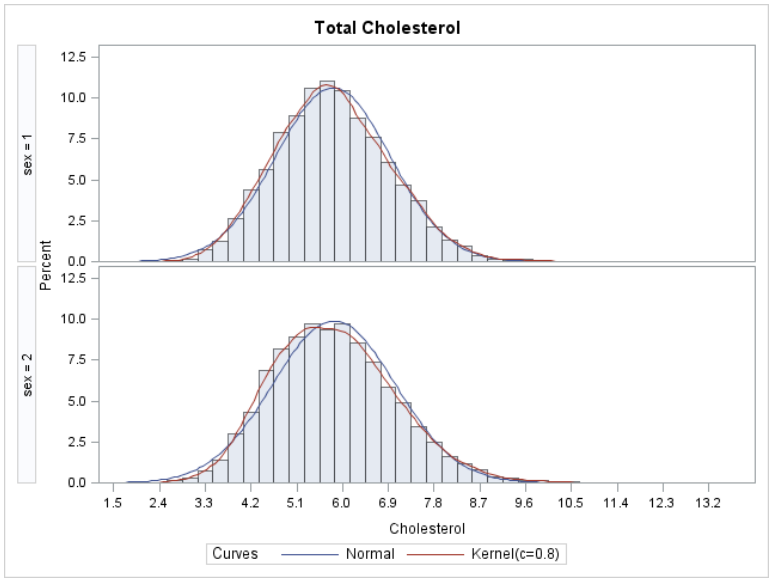
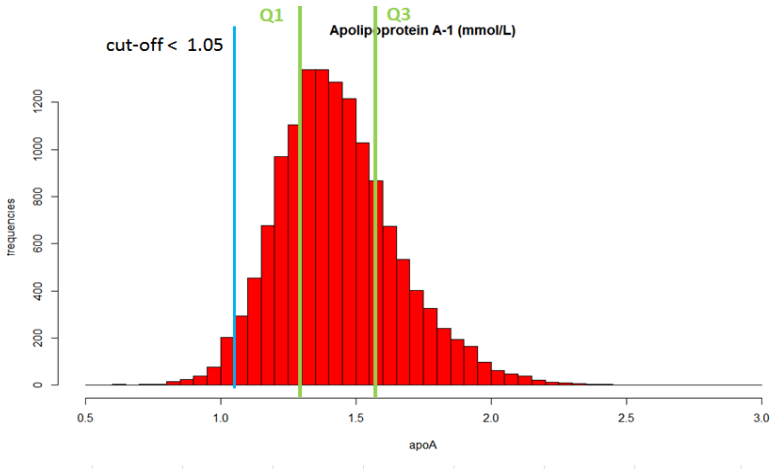


Figure 31. Histogram of the data distribution for ApoA. ApoA presents a normal distribution. A) The clinical cut-offs (blue line) represent less population than the quartiles (green lines). B) Histograms by sex status (1=male, 2=female).

A)



B)

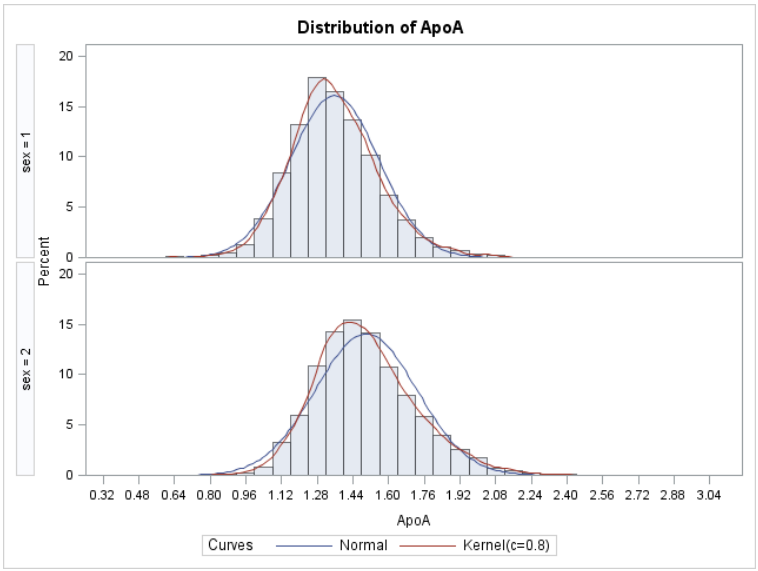
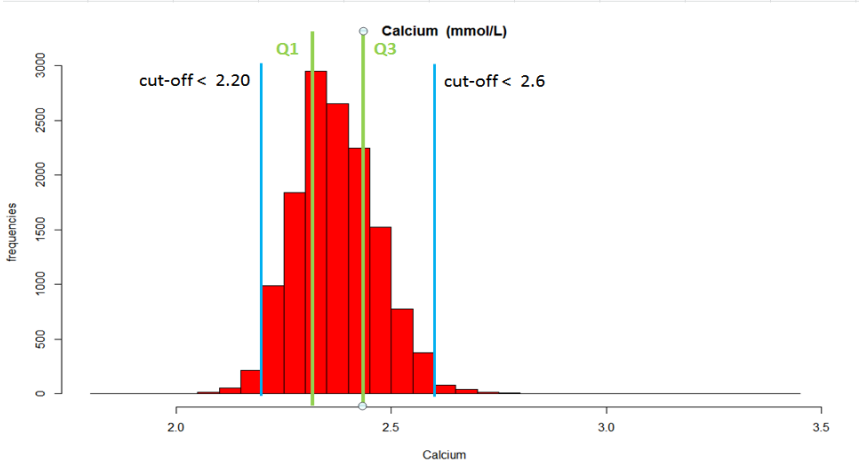


Figure 32. Histogram of the data distribution for Calcium. Calcium presents a normal distribution. A) The clinical cut-offs (blue lines) represent less population than the quartiles (green lines). B) Histograms by sex status (1=male, 2=female).

A)



B)

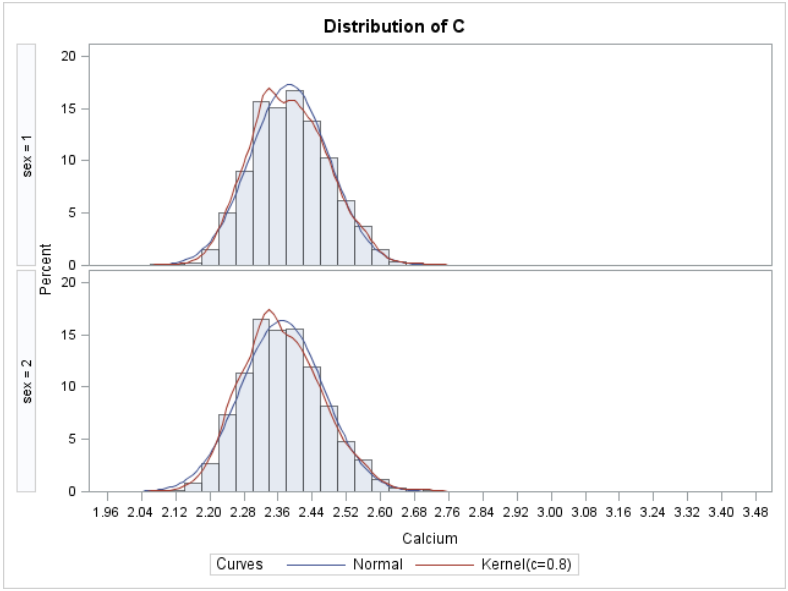
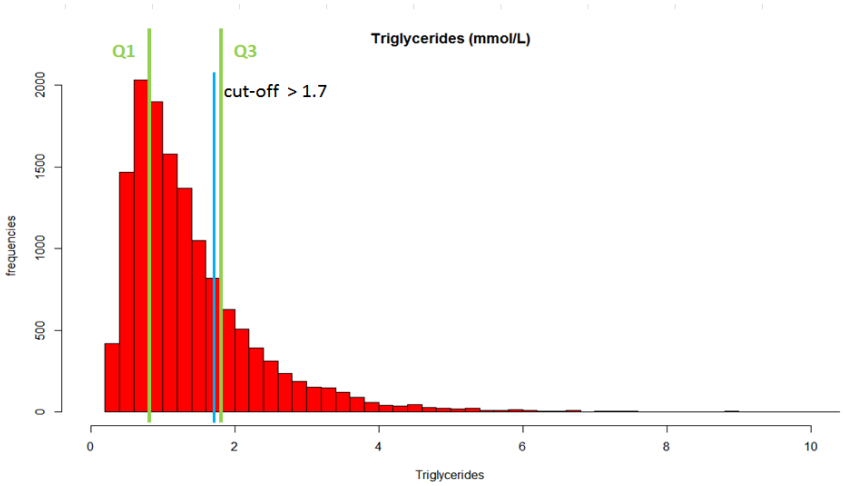


Figure 33. Histogram of the data distribution for TG. TG presents a skewed distribution. A) The blue line is the cut-off value, while green lines are the quartiles values. Approximately 30 % of the population has higher values for TC (cut-off >1.7 mmol/L). B) Histograms by sex status (1=male, 2=female).

A)



B)

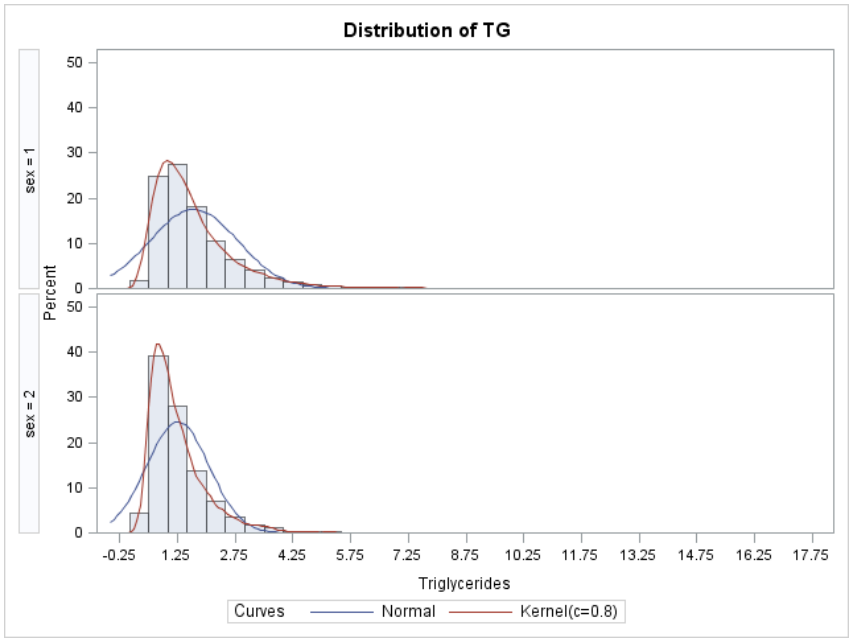


Figure 34. MODEL A. Line-graph depicting the goodness of fit indicators (AIC and BIC) for different LCA analysis. Figures A and B show the total population based on biomarkers based on clinical cut-offs run on R and run on SAS. Figures C and D show the training and the testing set based on biomarkers categorised based on clinical cut-offs run on R. The fit indicators decreased rapidly and stabilised after 3 to 4 classes in figures A, C and D (red arrow).

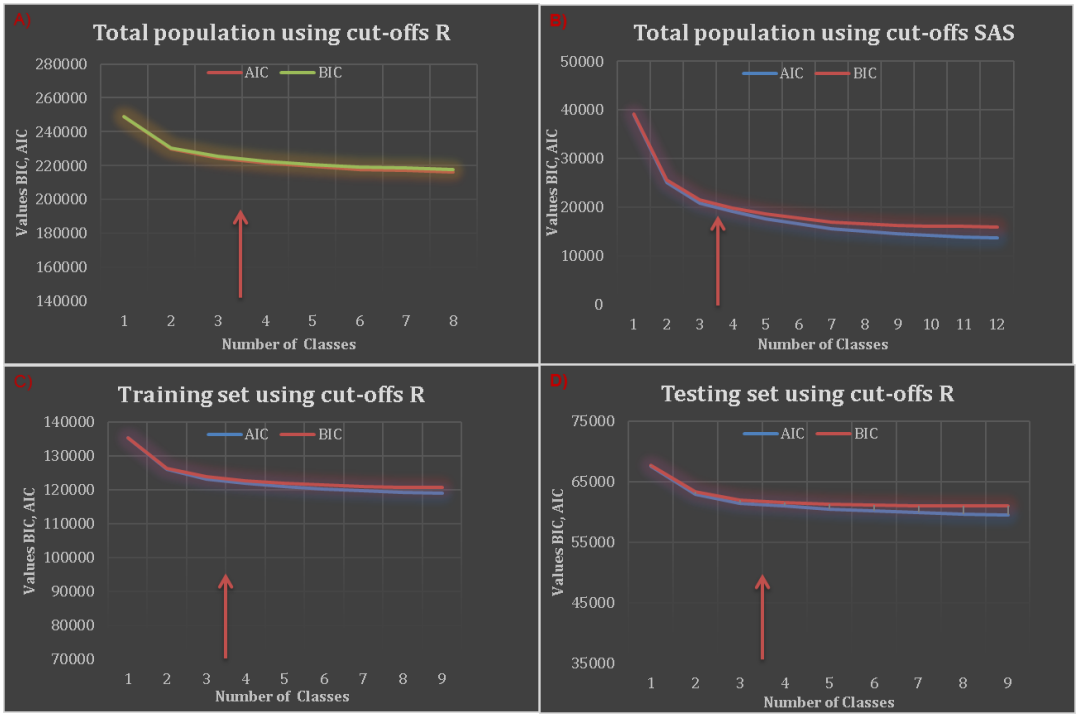


Figure 35. MODEL A. Line-graph depicting the goodness of fit indicators (X^2) for total population based on biomarkers based on clinical cut-offs run on R (same analysis as figure 33a). A minimum is reached at 3 classes (red arrow).

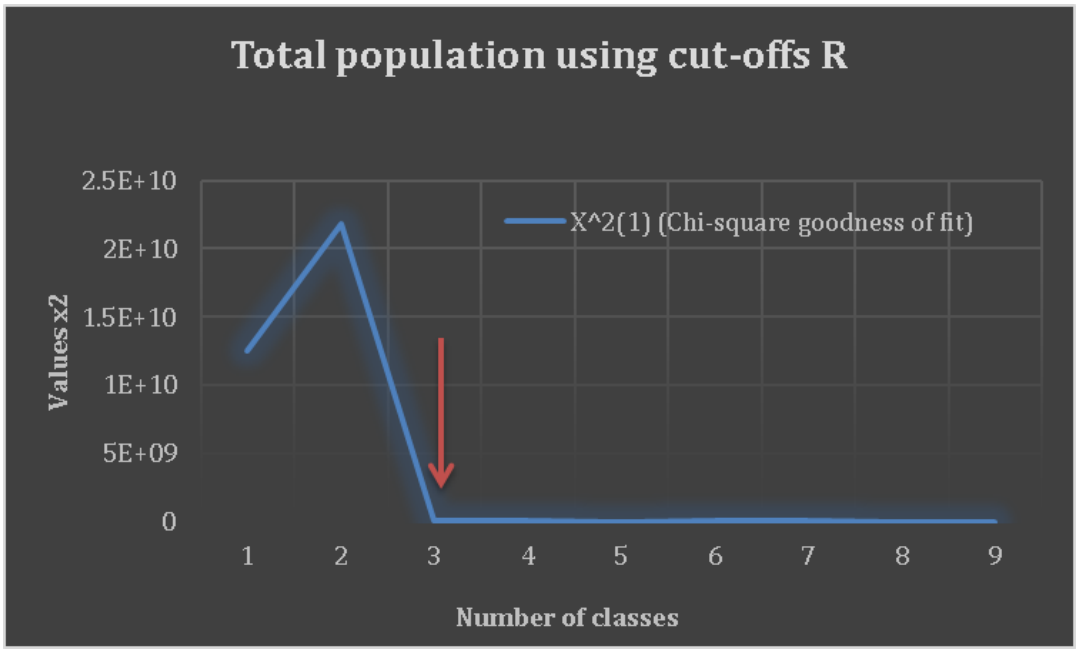


Table 8. MODEL A. Class membership probabilities for abnormal clinical values of serum markers for LCA two latent class model in the dataset. The numbers represent the probability of having an abnormal value for a biomarker in each class. * The clinical cut-offs used are identical to those defined in Table 5.

LCA-derived Classes	**Class 1**	**Class 2**
% on the population	36%	**64%**
Biological interpretation	Dyslipidaemia	Normal
Glucose ≥6.11 mmol/L	0.16	0.07
Fructosamine ≥ 2.60 mmol/L	0.06	0.02
Total Cholesterol ≥6.50 mmol/L	**0.67**	0.07
HDL < 1.03 mmol/L	0.24	0.03
LDL ≥4.10 mmol/L	**0.75**	0.07
Triglycerides ≥1.71 mmol/L	**0.50**	0.12
ApoA < 1.05 mmol/L	0.05	0.01
ApoB ≥ 1.50 mmol/L	**0.56**	0.00
ALT ≥ 50 IU/L	0.15	0.07
AST ≥ 45 IU/L	0.05	0.03
GGT 18-36 IU/L	0.40	0.35
GGT 36-72 IU/L	0.23	0.11
GGT ≥72 IU/L	0.12	0.05
Creatinine low µmol/L*	0.03	0.07
Creatinine high µmol/L*	0.27	0.15
Albumin <35 g/L	0.00	0.00
WBC ≥10^9 cells/L	0.07	0.04
CRP ≥10 mg/L	0.07	0.06
Iron low µmol/L*	0.04	0.07
Iron high µmol/L*	0.02	0.04
TIBC low mg/dL*	0.30	0.31
TIBC high mg/dL*	0.17	0.23
Phosphate low mmol/L*	0.02	0.01
Phosphate high mmol/L*	0.05	0.05
Calcium low mmol/L*	0.01	0.02
Calcium high mmol/L*	0.03	0.01

The following abbreviations have been used in Table 7: High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Table 9. MODEL A. Class membership probabilities for abnormal clinical values of serum markers for LCA three latent class model in the dataset. The numbers represent the probability of having an abnormal value for a biomarker in each class. * The clinical cut-offs used are identical to those defined in Table 5.

LCA-derived Classes	**Class 1**	**Class2**	**Class3**
% on the population	16%	29%	**55%**
Biological interpretation	Metabolic syndrome	Dyslipidaemia	Normal
Glucose ≥6.11 mmol/L	**0.25**	0.12	0.04
Fructosamine ≥ 2.60 mmol/L	0.08	0.04	0.01
Total Cholesterol ≥6.50 mmol/L	0.08	**0.86**	0.06
HDL < 1.03 mmol/L	**0.39**	0.13	0.00
LDL ≥4.10 mmol/L	0.08	**0.96**	0.07
Triglycerides ≥1.71 mmol/L	**0.56**	**0.42**	0.08
ApoA < 1.05 mmol/L	0.11	0.02	0.00
ApoB ≥ 1.50 mmol/L	0.15	**0.62**	0.00
ALT ≥ 50 IU/L	**0.34**	0.11	0.01
AST ≥ 45 IU/L	0.13	0.03	0.01
GGT 18-36 IU/L	0.37	**0.40**	0.35
GGT 36-72 IU/L	0.29	0.20	0.09
GGT ≥72 IU/L	0.21	0.10	0.02
Creatinine low µmol/L*	0.02	0.04	0.07
Creatinine high µmol/L*	**0.34**	0.23	0.13
Albumin <35 g/L	0.00	0.00	0.00
WBC ≥10^9 cells/L	0.07	0.06	0.04
CRP ≥10 mg/L	0.10	0.06	0.05
Iron low µmol/L*	0.05	0.03	0.08
Iron high µmol/L*	0.05	0.02	0.04
TIBC low mg/dL*	0.30	0.29	0.31
TIBC high mg/dL*	0.21	0.17	0.23
Phosphate low mmol/L*	0.02	0.01	0.01
Phosphate high mmol/L*	0.07	0.05	0.05
Calcium low mmol/L*	0.02	0.01	0.02
Calcium high mmol/L*	0.02	0.03	0.01

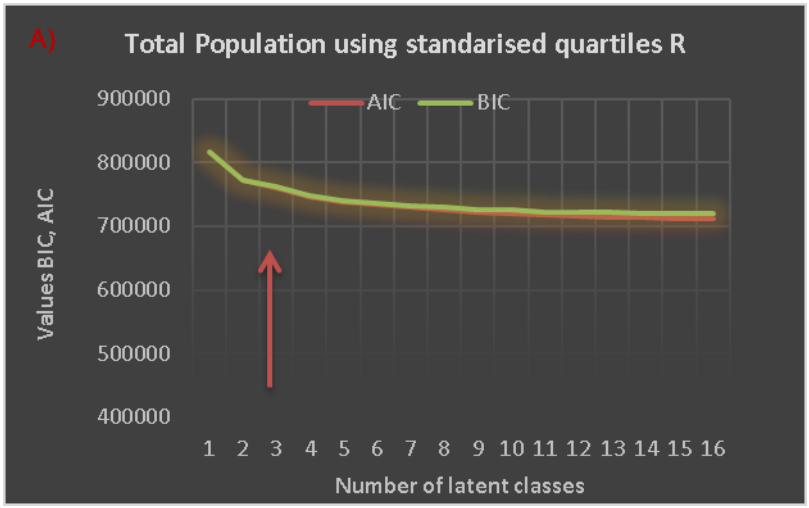
The following abbreviations have been used in Table 8: High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Table 10. MODEL A. Class membership probabilities for abnormal clinical values of serum markers for LCA four latent class model in the dataset. The numbers represent the probability of having an abnormal value for a biomarker in each class. * The clinical cut-offs used are identical to those defined in Table 5.

LCA-derived Classes	**Class 1**	**Class2**	**Class3**	**Class 4**
% on the population	12%	19%	19%	**50%**
Biological interpretation	Dyslipidaemia	Dyslipidaemia	Dyslipidaemia	Normal
Glucose ≥6.11 mmol/L	0.23	0.06	0.23	0.04
Fructosamine ≥ 2.60 mmol/L	0.09	0.01	0.07	0.01
Total Cholesterol ≥6.50 mmol/L	**0.74**	**0.83**	0.08	0.04
HDL < 1.03 mmol/L	0.39	0.00	0.29	0.00
LDL ≥4.10 mmol/L	**0.77**	**0.97**	0.09	0.03
Triglycerides ≥1.71 mmol/L	**0.82**	0.17	**0.45**	0.07
ApoA < 1.05 mmol/L	0.04	0.00	0.10	0.00
ApoB ≥ 1.50 mmol/L	**0.92**	**0.42**	0.02	0.00
ALT ≥ 50 IU/L	0.23	0.03	0.32	0.00
AST ≥ 45 IU/L	0.06	0.00	0.12	0.00
GGT 18-36 IU/L	0.39	**0.40**	0.38	0.34
GGT 36-72 IU/L	0.30	0.14	0.28	0.07
GGT ≥72 IU/L	0.19	0.04	0.19	0.01
Creatinine low µmol/L*	0.03	0.04	0.02	0.07
Creatinine high µmol/L*	0.32	0.17	0.33	0.12
Albumin <35 g/L	0.00	0.00	0.00	0.00
WBC $\geq 10^9$ cells/L	0.08	0.05	0.07	0.03
CRP ≥10 mg/L	0.07	0.05	0.11	0.05
Iron low µmol/L*	0.03	0.04	0.05	0.08
Iron high µmol/L*	0.02	0.02	0.05	0.04
TIBC low mg/dL*	0.32	0.29	0.29	0.32
TIBC high mg/dL*	0.14	0.20	0.22	0.23
Phosphate low mmol/L*	0.02	0.01	0.02	0.01
Phosphate high mmol/L*	0.06	0.03	0.07	0.05
Calcium low mmol/L*	0.01	0.01	0.02	0.02
Calcium high mmol/L*	0.04	0.02	0.02	0.01

The following abbreviations have been used in Table 9: High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Figure 36. MODEL B. Line-graph depicting the goodness of fit indicators (AIC, BIC (A) and X^2 (B)) for LCA of the total population based on biomarkers categorised based on standardised quartiles run on R. The minimum is not clearly reached this time for any of the indicators.



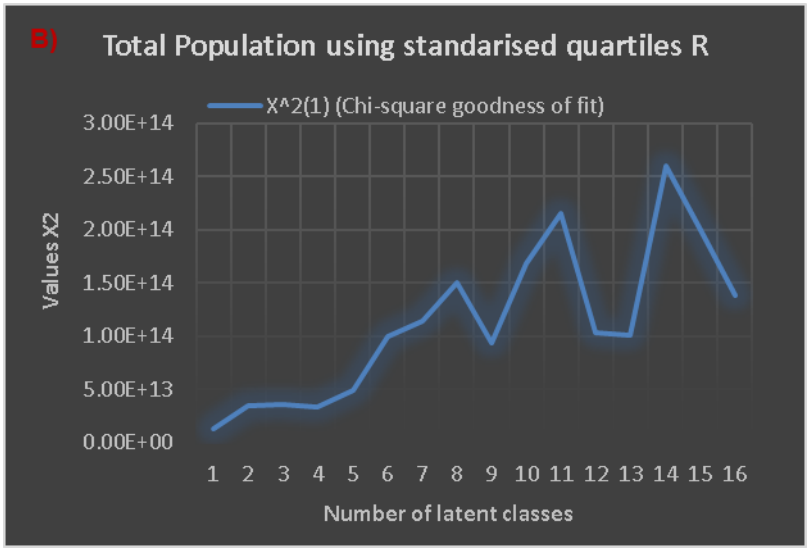


Table 11. MODEL B. Class membership probabilities of the serum markers in a LCA for 2 latent classes for standardised quartiles values of the biomarkers run in R. The numbers represent the probability of having an abnormal value for a biomarker in each class. Marked quartiles in red represent clinical abnormal values.

LCA-derived Classes	**Class 1**	**Class 2**
% on the population	51%	49%
Biological interpretation	High lipids profile	Low lipids profile
Glucose 1st quartile	0.18	**0.33**
Glucose 2nd quartile	0.28	0.34
Glucose 3rd quartile	0.21	0.17
Glucose 4th quartile	**0.33**	0.16
Total Cholesterol 1st quartile	0.07	**0.47**
Total Cholesterol 2nd quartile	0.14	0.25
Total Cholesterol 3rd quartile	0.34	0.22
Total Cholesterol 4th quartile	**0.45**	0.06
HDL 1st quartile	**0.42**	0.07
HDL 2nd quartile	0.3	0.2
HDL 3rd quartile	0.19	0.32
HDL 4th quartile	0.1	**0.41**
LDL 1st quartile	0.05	**0.46**
LDL 2nd quartile	0.16	0.34
LDL 3rd quartile	0.32	0.18
LDL 4th quartile	**0.46**	0.03
Triglycerides 1st quartile	0.09	**0.5**
Triglycerides 2nd quartile	0.21	0.29
Triglycerides 3rd quartile	0.26	0.14
Triglycerides 4th quartile	**0.44**	0.06
ApoA 1st quartile	**0.36**	0.15
ApoA 2nd quartile	0.28	0.22
ApoA 3rd quartile	0.21	0.25
ApoA 4th quartile	0.14	**0.37**
ApoB 1st quartile	0.01	**0.5**
ApoB 2nd quartile	0.15	0.36
ApoB 3rd quartile	0.34	0.13
ApoB 4th quartile	**0.49**	0.01

The following abbreviations have been used in Table 10: High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Apolipoprotein A-1 (ApoA) and Apolipoprotein B (ApoB).

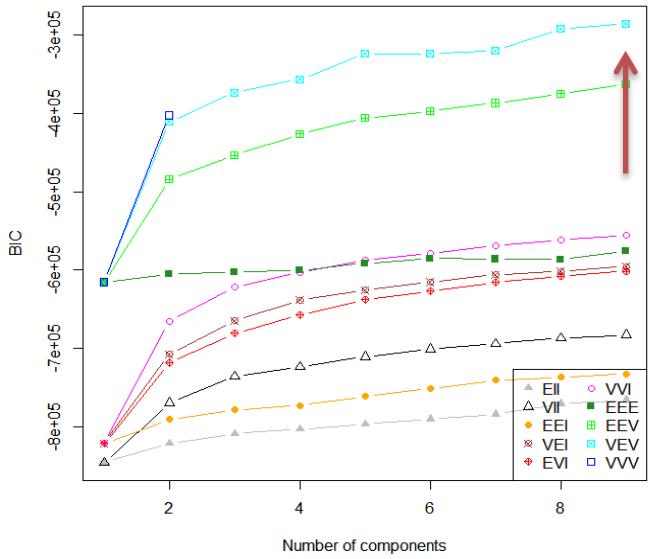
Table 12. MODEL B. Class membership probabilities of the serum markers in a LCA for 3 latent classes for standardised quartiles values of the biomarkers run in R. The numbers represent the probability of having an abnormal value for a biomarker in each class. Marked quartiles in red represent clinical abnormal values.

LCA-derived Classes	**Class 1**	**Class2**	**Class3**
% on the population	34%	37%	28%
Biological interpretation	High lipids profile	Low lipids profile	Normal
Glucose 1st quartile	0.17	**0.34**	0.23
Glucose 2nd quartile	0.28	0.34	0.3
Glucose 3rd quartile	0.21	0.17	0.19
Glucose 4th quartile	**0.33**	0.15	0.28
TC 1st quartile	0.01	**0.49**	0.3
TC 2nd quartile	0.05	0.22	0.35
TC 3rd quartile	0.29	0.23	0.34
TC 4th quartile	**0.66**	0.07	0.01
HDL 1st quartile	**0.39**	0.02	0.38
HDL 2nd quartile	0.25	0.14	**0.4**
HDL 3rd quartile	0.21	0.31	0.22
HDL 4th quartile	0.15	**0.52**	0
LDL 1st quartile	0.01	**0.51**	0.19
LDL 2nd quartile	0.05	0.31	**0.41**
LDL 3rd quartile	0.26	0.16	0.36
LDL 4th quartile	**0.67**	0.03	0.03
TG 1st quartile	0.08	**0.55**	0.2
TG 2nd quartile	0.19	0.28	0.28
TG 3rd quartile	0.25	0.12	0.26
TG 4th quartile	**0.49**	0.05	0.25
ApoA 1st quartile	0.29	0.09	**0.46**
ApoA 2nd quartile	0.26	0.19	0.34
ApoA 3rd quartile	0.23	0.25	0.19
ApoA 4th quartile	0.22	**0.47**	0.01
ApoB 1st quartile	0	**0.56**	0.14
ApoB 2nd quartile	0.02	0.33	**0.45**
ApoB 3rd quartile	0.25	0.11	**0.4**
ApoB 4th quartile	**0.73**	0.01	0.01

The following abbreviations have been used in Table 11: Total Cholesterol (TC), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Triglycerides (TG), Apolipoprotein A-1 (ApoA) and Apolipoprotein B (ApoB).

Figure 37. MODEL C. Line-graph depicting the goodness of fit indicator BIC for LCA of the total population based on standardised continuous data (A) and continuous data (B) on MCLUST R. 9 classes indicated the best model for *standardised* continuous data (red arrow) while continuous show a BIC maximum with only one class (red arrow). This LCA implementation on MCLUST allowed for testing of multiple different data distribution models represented by the letters in the box on the bottom right (e.g.: EII, VII, EEI, etc.)

A)



B)

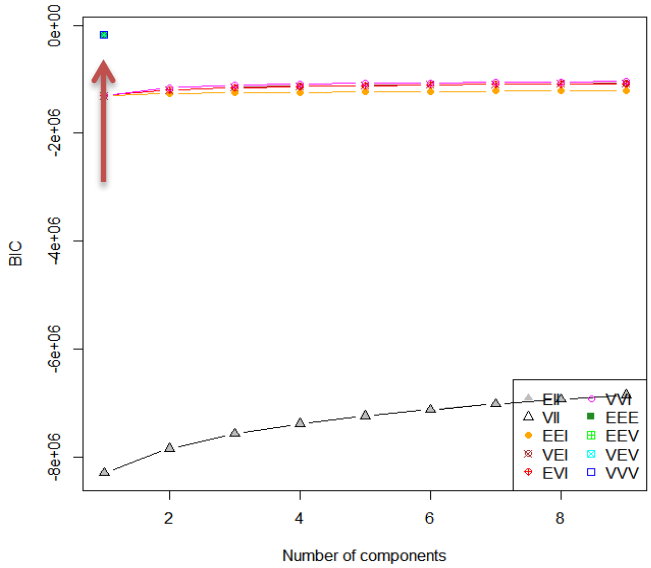


Figure 38. Line-graph depicting the goodness of fit indicator X^2 for LCA of the total population using biomarkers dichotomised based on clinical cut-offs run on R for MODEL D (A), MODEL E (B), MODEL F (C) and MODEL G (D). A minimum is indicated with a red arrow.

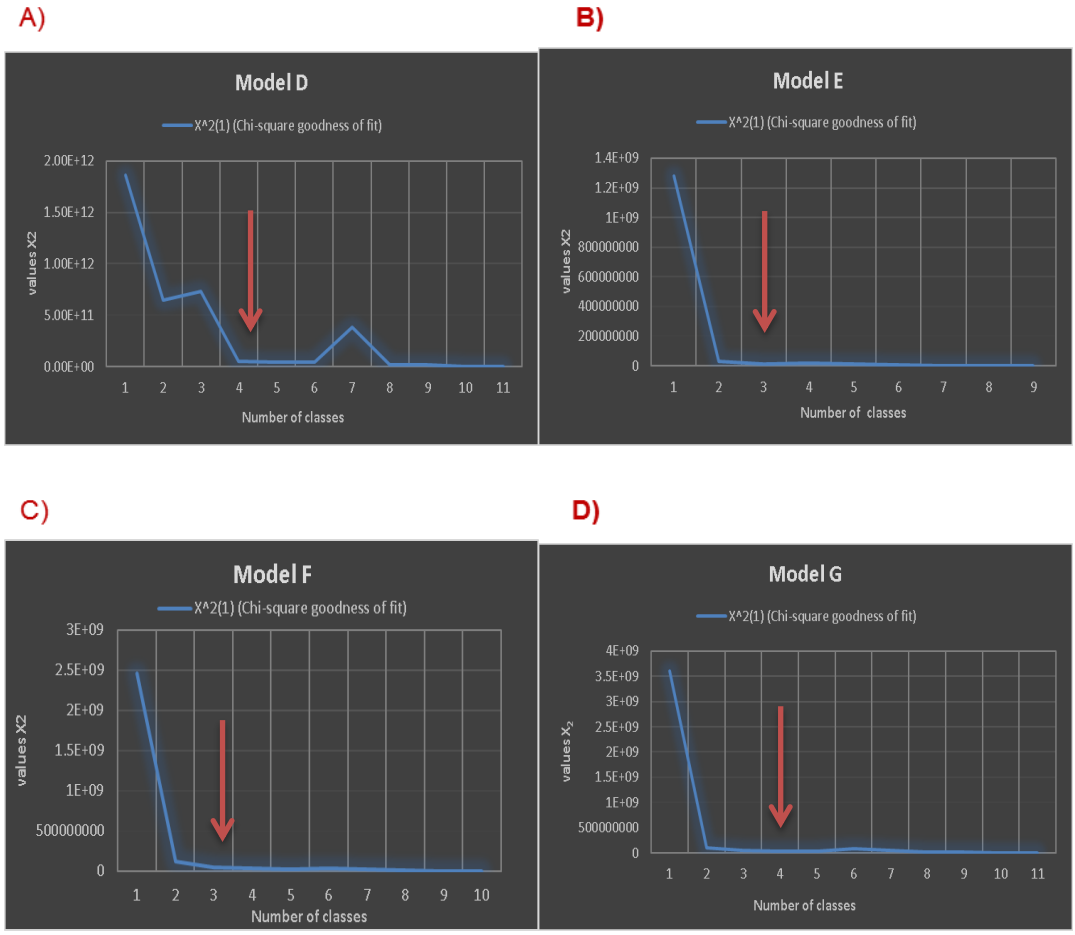


Table 13. MODEL D. Class membership probabilities of the serum markers in a LCA for 4 latent classes.
The numbers represent the probability of having an abnormal value for a biomarker in each class.
* The clinical cut-offs used are identical to those defined in Table 5.

LCA-derived Classes	**class 1**	**class 2**	**class 3**	**Class 4**
% on the population	12%	16%	20%	52%
Biological interpretation	Liver	Dyslipidaemia	Dyslipidaemia	Normal
Glucose ≥6.11 mmol/L	0.2272	0.2089	0.0735	0.0467
Fructosamine ≥ 2.60 mmol/L	0.0644	0.0799	0.0184	0.0128
TC ≥6.50 mmol/L	0.1238	0.5482	**0.8235**	0.038
HDL < 1.03 mmol/L	0.1039	**0.5498**	0	0.0075
LDL ≥4.10 mmol/L	0.0607	0.6104	**0.9711**	0.0412
Triglycerides ≥1.71 mmol/L	0.5569	**0.7996**	0.1625	0.0409
ApoA-I < 1.05 mmol/L	0.0221	0.1101	0	0.0064
ApoB ≥ 1.50 mmol/L	0.0715	**0.6088**	0.4719	0.0016
ApoB/ApoA-I ≥ 1.00	0.1636	**0.9231**	**0.5375**	0.0424
Total cholesterol/HDL ≥ 5.00	0.1122	**0.9844**	0.1732	0.001
LDL/HDL ≥ 3.50	0.0056	**0.8853**	0.2013	0.0005
Log (triglycerides/HDL) ≥ 0.5	0.3553	**0.7832**	0.0158	0
ALT ≥ 50 IU/L	**0.3836**	0.1865	0.055	0.0097
AST ≥ 45 IU/L	0.1614	0.0474	0.0128	0.0042
GGT Elevated36-72 IU/L	**0.285**	0.2616	0.1611	0.0835
GGT Highly elevated ≥72 IU/L	0.2391	0.1447	0.0573	0.0172
Creatinine low µmol/L*	0.0062	0.0023	0	0.0034
Creatinine high µmol/L*	0.1262	0.1721	0.1224	0.08
Albumin <35 g/L	0.0025	0.0049	0.0016	0.0012
WBC ≥10^9 cells/L	0.062	0.0832	0.0481	0.034
CRP ≥10 mg/L	0.1167	0.0878	0.043	0.0461
Iron low µmol/L*	0.0575	0.0528	0.024	0.0504
Iron high µmol/L*	0.0635	0.0188	0.0199	0.037
TIBC low mg/dL*	0.311	**0.3388**	0.2599	0.2973
TIBC high mg/dL*	0.2649	0.1722	0.1923	0.2198

Phosphate low mmol/L*	0.0068	0.0056	0.0053	0.0081
Phosphate high mmol/L*	0.0845	0.0654	0.0342	0.0482
Calcium low mmol/L*	0.016	0.0122	0.0076	0.0165
Calcium high mmol/L*	0.025	0.0254	0.0193	0.0111

The following abbreviations have been used in Table 12: Total Cholesterol (TC), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Table 14. Hazard ratios and 95 % confidence interval for the association of LCA-derived metabolic classes and overall cancer risk crude and adjusted analysis using calendar-time as a time scale for Model D, Model E, Model F and Model G. Marked in red classes that showed a statistical significance association.

	HR (95% CI) crude	**HR (95% CI) adjusted for Age, Sex, CCI and Education Status**
MODEL D: All Lipids		
Normal (52%)	1.00 (ref)	1.00 (ref)
Dyslipidaemia (20%)	1.44 (1.27 - 1.63)	1.10 (0.97 -1.26)
Dyslipidaemia (16%)	1.49 (1.33 - 1.67)	1.02 (0.90 - 1.14)
Liver (12%)	1.40 (1.25 - 1.61)	1.19 (1.02 - 1.37)
MODEL E: No Lipids		
Normal (78%)	1.00 (ref)	1.00 (ref)
Liver (15%)	1.15 (1.02 - 1.30)	1.06 (0.93 - 1.21)
Iron (7%)	1.11 (0.93 - 1.32)	1.09 (0.91 - 1.30)
MODEL F: TG TC		
Normal (68%)	1.00 (ref)	1.00 (ref)
Liver & Lipids (25%)	1.32 (1.19 - 1.45)	1.16 (1.04 - 1.28)
Iron (7%)	1.19 (0.99 - 1.42)	1.17 (0.97 - 1.40)
MODEL G: Lipid ratios		
Normal (63%)	1.00 (ref)	1.00 (ref)
Lipids (23%)	1.36 (1.22- 1.51)	1.10 (0.98 - 1.23)
Liver (9%)	1.26 (1.08 -1.47)	1.29 (1.10 - 1.52)
Iron (6%)	1.20 (0.99 - 1.44)	1.16 (0.96 - 1.41)

Table 15. MODEL E. Class membership probabilities of the serum markers in a LCA for 3 latent classes. The numbers represent the probability of having an abnormal value for a biomarker in each class. *The clinical cut-offs used are identical to those defined in Table 5.

LCA-derived Classes	**class 1**	**class 2**	**class 3**
% on the population	7%	78%	15%
Biological interpretation	Low iron	Normal	Liver
Glucose ≥6.11 mmol/L	0.1221	0.0506	0.2938
Fructosamine ≥ 2.60 mmol/L	0.0409	0.0112	0.1066
ALT ≥ 50 IU/L	0.0476	0	**0.4989**
AST ≥ 45 IU/L	0.0223	0.0028	0.1603
GGT Elevated36-72 IU/L	0.193	0.0956	**0.3627**
GGT Highly elevated ≥72 IU/L	0.0737	0.0132	**0.3171**
Creatinine low µmol/L*	0.005	0.0023	0.0043
Creatinine high µmol/L*	0.1171	0.1007	0.1386
Albumin <35 g/L	0.0103	0.0012	0.0015
WBC ≥10^9 cells/L	0.1436	0.0349	0.054
CRP ≥10 mg/L	0.2611	0.0337	0.0724
Iron low µmol/L*	**0.4953**	0	0.007
Iron high µmol/L*	0	0.032	0.0601
TIBC low mg/dL*	**0.9974**	0.2271	0.2325
TIBC high mg/dL*	0.0007	0.2212	0.2852
Phosphate low mmol/L*	0.008	0.0069	0.0069
Phosphate high mmol/L*	0.1122	0.0454	0.0541
Calcium low mmol/L*	0.0418	0.0112	0.0114
Calcium high mmol/L*	0.0143	0.0138	0.0297

The following abbreviations have been used in Table 14: Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Table 16. MODEL F. Class membership probabilities of the serum markers in a LCA for 3 latent classes. The numbers represent the probability of having an abnormal value for a biomarker in each class. *The clinical cut-offs used are identical to those defined in Table 5.

LCA-derived Classes	**class 1**	**class 2**	**class 3**
% on the population	68%	25%	7%
Biological interpretation	Normal	Liver and Lipids	Iron
Glucose ≥6.11 mmol/L	0.0344	0.2592	0.1095
Fructosamine ≥ 2.60 mmol/L	0.007	0.0878	0.0368
Total cholesterol ≥6.50 mmol/L	0.231	**0.4314**	0.1862
Triglycerides ≥1.71 mmol/L	0.1354	**0.5477**	0.2289
AST ≥ 45 IU/L	0.0024	0.109	0.0288
ALT ≥ 50 IU/L	0	**0.3348**	0.0623
GGT Elevated36-72 IU/L	0.0677	**0.3476**	0.1838
GGT Highly elevated ≥72 IU/L	0.0053	0.2389	0.0751
Creatinine low µmol/L*	0.0024	0.0035	0.0052
Creatinine high µmol/L*	0.0935	0.144	0.116
Albumin <35 g/L	0.001	0.0016	0.0114
WBC ≥10^9 cells/L	0.0313	0.0595	0.143
CRP ≥10 mg/L	0.03	0.0707	0.2749
Iron low µmol/L*	0.0002	0.0025	**0.5467**
Iron high µmol/L*	0.033	0.0479	0
TIBC low mg/dL*	0.2312	0.2425	**0.9982**
TIBC high mg/dL*	0.2242	0.2515	0
Phosphate low mmol/L*	0.0067	0.0072	0.0086
Phosphate high mmol/L*	0.0433	0.0585	0.1112
Calcium low mmol/L*	0.012	0.0089	0.0461
Calcium high mmol/L*	0.012	0.0293	0.0128

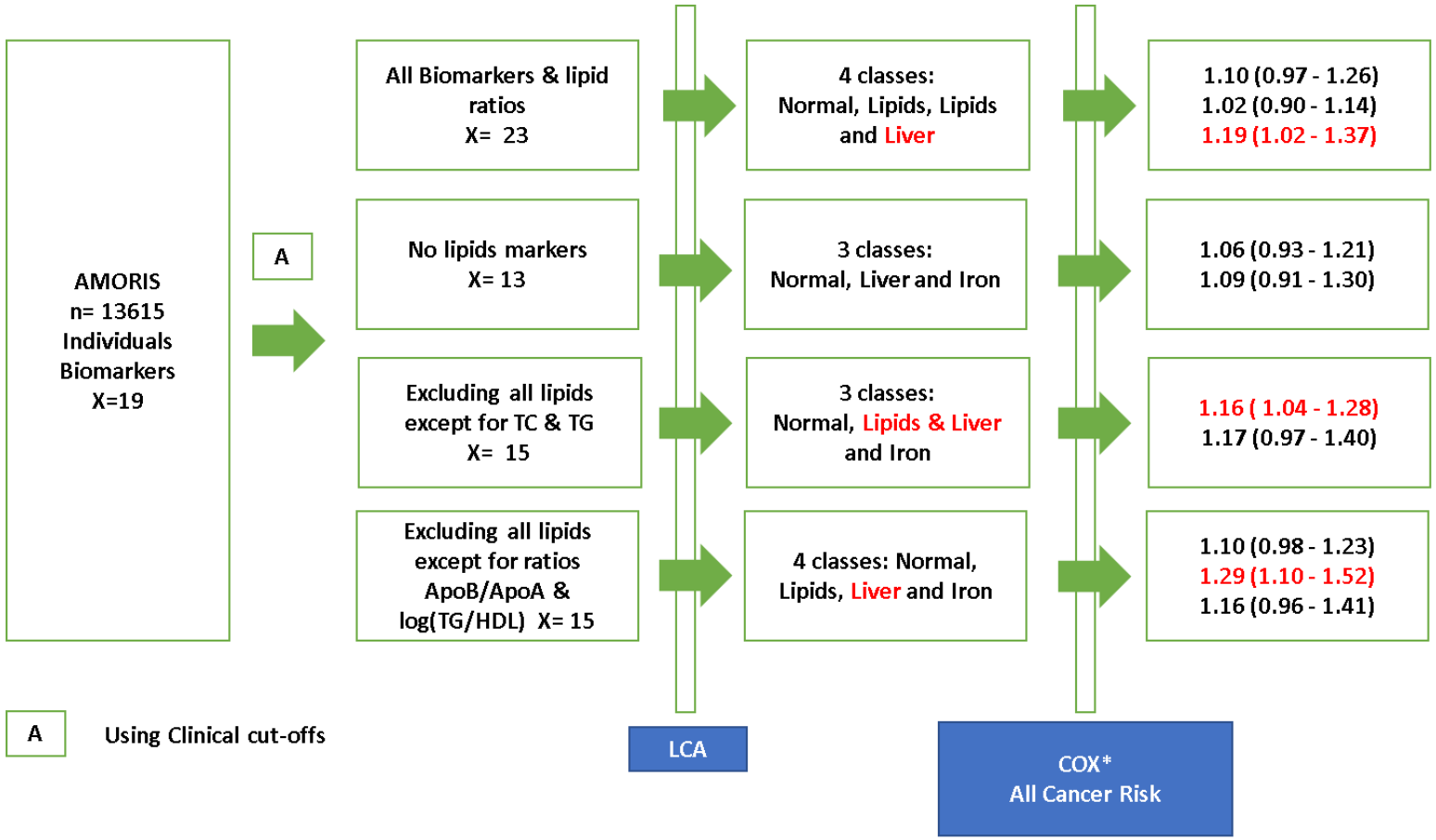
The following abbreviations have been used in Table 15: Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Table 17. MODEL G. Class membership probabilities of the serum markers in a LCA for 4 latent classes. The numbers represent the probability of having an abnormal value for a biomarker in each class. *The clinical cut-offs used are identical to those defined in Table 5.

LCA-derived Classes	**Class 1**	**Class 2**	**Class 3**	**Class 4**
% on the population	63%	23%	9%	6%
Biological interpretation	Normal	Lipids	Liver	Iron
Glucose ≥6.11 mmol/L	0.0342	0.2401	0.2174	0.0919
Fructosamine ≥ 2.60 mmol/L	0.0039	0.0967	0.0555	0.028
ApoB/ApoA-I ≥ 1.00	0.132	**0.684**	0.4519	0.248
Log (Triglycerides/HDL) ≥ 0.50	0.0126	**0.5436**	0.3852	0.1421
AST ≥ 45 IU/L	0.0052	0.0045	**0.3168**	0.018
ALT ≥ 50 IU/L	0.0051	0.0107	**1**	0.0291
GGT Elevated36-72 IU/L	0.0848	0.2532	**0.3521**	0.1732
GGT Highly elevated ≥72 IU/L	0.024	0.0843	**0.4098**	0.0619
Creatinine low µmol/L*	0.0022	0.0037	0.0041	0.0051
Creatinine high µmol/L*	0.0822	0.1765	0.1166	0.1116
Albumin <35 g/L	0.0007	0.0022	0.0024	0.0114
WBC ≥10^9 cells/L	0.0265	0.0786	0.0438	0.1344
CRP ≥ 10 mg/L	0.0282	0.0715	0.0771	0.274
Iron low µmol/L*	0.0001	0.004	0.0281	**0.5527**
Iron high µmol/L*	0.0404	0.0155	0.0712	0
TIBC low mg/dL*	0.2201	0.2807	0.2622	**1**
TIBC high mg/dL*	0.2438	0.1707	0.2984	0
Phosphate low mmol/L*	0.0078	0.0041	0.0063	0.0098
Phosphate high mmol/L*	0.0425	0.0611	0.0544	0.111
Calcium low mmol/L*	0.0124	0.0092	0.0099	0.0458
Calcium high mmol/L*	0.0121	0.0253	0.0299	0.0135

The following abbreviations have been used in Table 16: High Density Lipoprotein (HDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Figure 39. Schematic description of the sensitivity analysis of the panel of biomarkers for the statistical pipeline (MODEL D, MODEL E, MODEL F and MODEL G).



* Time scale adjusted for age, sex, CCI and education status

Figure 40. Line-graph depicting the goodness of fit indicators AIC, BIC (A) and X^2 (B) for LCA of the total population using biomarkers dichotomised based on clinical cut-offs run on R for MODEL H. A minimum is indicated with a red arrow.

A)

B)

.

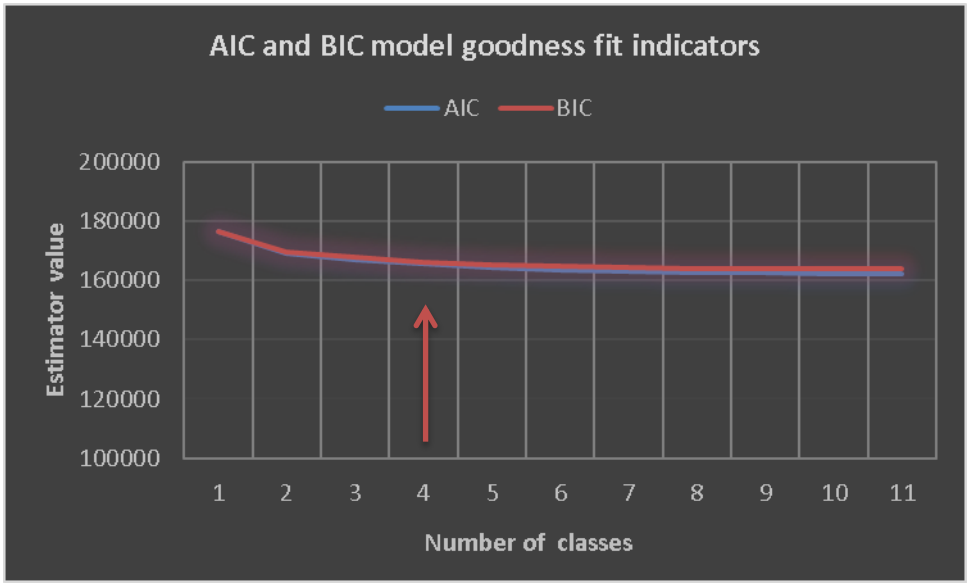


Table 18. MODEL H. Predicted class memberships of the clinically abnormal biomarkers cut-off values for the estimated class population shares for the four different LCA classes. The numbers represent the probability of having an abnormal value for a biomarker in each class. *The clinical cut-offs used are identical to those defined in Table 5.

LCA-derived Classes	**Class 1**	**Class 2**	**Class 3**	**Class 4**
% on the population	63%	22%	9%	6%
Biological interpretation	Normal	Lipids	Liver	Iron/ Inflammation
Glucose ≥6.11 mmol/L	0.0342	0.2401	0.2174	0.0919
Fructosamine ≥ 2.60 mmol/L	0.0039	0.0967	0.0555	0.028
ApoB/ApoA-I ≥ 1.00	0.132	**0.684**	0.4519	0.248
Log (Triglycerides/HDL) ≥ 0.50	0.0126	**0.5436**	0.3852	0.1421
ALT ≥ 50 IU/L	0.0051	0.0107	**1**	0.0291
AST ≥ 45 IU/L	0.0052	0.0045	**0.3168**	0.018
GGT Elevated36-72 IU/L	0.0848	0.2532	**0.3521**	0.1732
GGT Highly elevated ≥72 IU/L	0.024	0.0843	**0.4098**	0.0619
Creatinine low µmol/L*	0.0022	0.0037	0.0041	0.0051
Creatinine high µmol/L*	0.0822	0.1765	0.1166	0.1116
Albumin <35 g/L	0.0007	0.0022	0.0024	0.0114
WBC ≥ 10^9 cells/L	0.0265	0.0786	0.0438	0.1344
CRP ≥10 mg/L	0.0282	0.0715	0.0771	**0.274**
Iron low µmol/L*	0.0001	0.004	0.0281	**0.5527**
Iron high µmol/L*	0.0404	0.0155	0.0712	0
TIBC low mg/dL*	0.2201	0.2807	0.2622	**1**
TIBC high mg/dL*	0.2438	0.1707	0.2984	0
Phosphate low mmol/L*	0.0078	0.0041	0.0063	0.0098
Phosphate high mmol/L*	0.0425	0.0611	0.0544	0.111
Calcium low mmol/L*	0.0124	0.0092	0.0099	0.0458
Calcium high mmol/L*	0.0121	0.0253	0.0299	0.0135

The following abbreviations have been used in Table 16: High Density Lipoprotein (HDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Gamma-Glutamyl transferase (GGT), Leukocytes (WBC), C-Reactive Protein (CRP) and Total iron binding capacity (TIBC).

Figure 41. MODEL H. Class Membership Probabilities for abnormal clinical values of the serum markers for the four LCA – derived metabolic classes. The four different biomarker profiles are represented in the graph. This figure was included to facilitate the visualisation of the metabolic profiles described in Table 18.

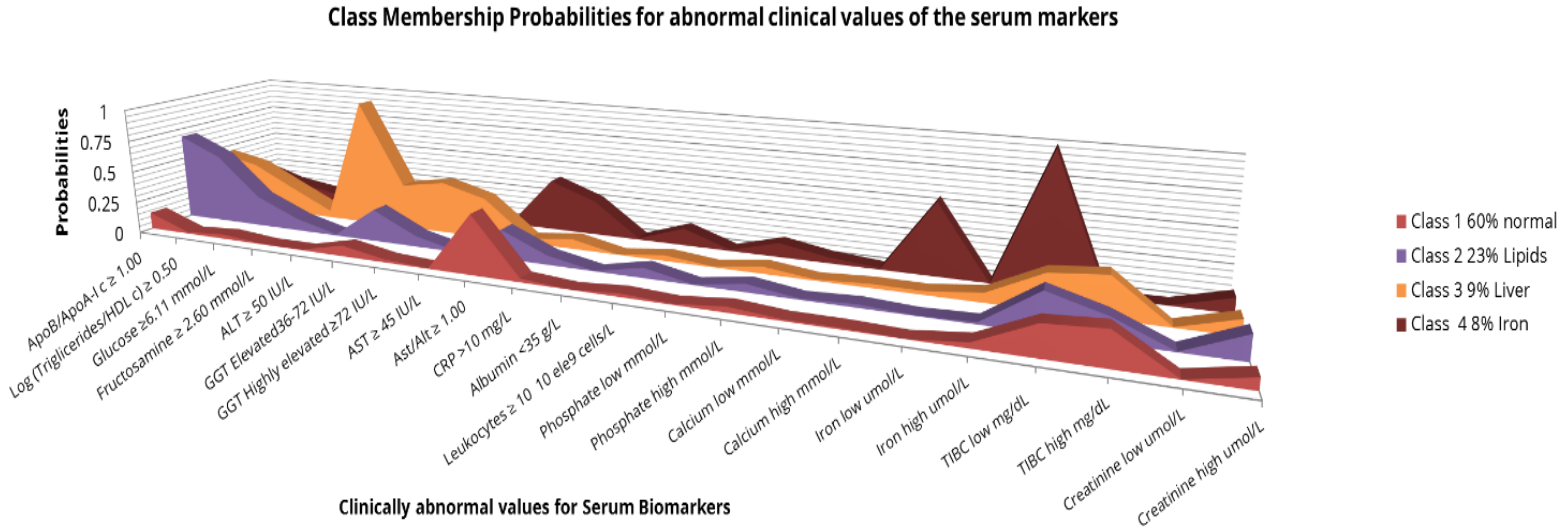


Table 19. Characteristics of the study population by LCA-derived metabolic classes based on MODEL H. All the serum markers are dichotomized using the standardized clinical cut-offs. Clinically abnormal cut-off values are highlighted for each biomarker. *The clinical cut-offs used are identical to those defined in Table 5.

	Class 1 Normal N=8612 (63.26%)	**Class 2 Lipids N=2936 (21.56%)**	**Class 3 Liver N=1262 (9.27%)**	**Class 4 Iron/Inflammation N=805 (5.91%)**
Age (years)				
Mean(SD)	51.25 (15.46)	55.17 (13.04)	48.98 (12.34)	51.71 (15.21)
Under 40	2126 (24.69)	360 (12.26)	295 (23.38)	170 (21.12)
40-50	2192 (25.45)	699 (23.81)	415 (32.88)	244 (30.31)
50-60	1779 (20.66)	813 (27.69)	313 (24.80)	160 (19.88)
Above 60	2515 (29.20)	1064 (36.24)	239 (18.94)	231 (28.70)
Sex				
Female	5742 (66.67)	1076 (36.65)	320 (25.36)	450 (55.90)
Male	2870 (33.33)	1860 (63.35)	942 (74.64)	355 (44.10)
Socio-economics Status				
High	3959 (45.97)	1515 (51.60)	672 (53.25)	347 (43.11)
Low	3257 (37.82)	972 (33.11)	466 (36.93)	312 (38.76)
Not employed or missing	1396 (16.21)	449 (15.29)	124 (9.83)	146 (18.14)
Educational Status				
High	2846 (34.91)	840 (30.26)	388 (32.17)	239 (31.04)
Middle	3525 (43.24)	1156 (41.64)	498 (41.29)	316 (41.04)
Low	1782 (21.86)	780 (28.10)	320 (26.53)	215 (27.92)
Missing +	459 (5.33)	160 (5.45)	56 (4.44)	35(4.35)
CCI				
0	7959 (92.42)	2497 (85.05)	1113 (88.19)	689 (85.59)
1	495 (5.75)	294 (10.01)	103 (8.16)	71 (8.82)
2	101 (1.17)	71 (0.52)	24 (1.90)	25 (3.11)
3+	57 (0.66)	74 (0.54)	22 (1.74)	20 (2.48)
Glucose (mmol/L)				
Mean(SD)	4.89 (0.75)	5.96 (2.29)	5.80 (2.18)	5.26 (1.87)
< 6.11	8316 (96.56)	2188 (74.52)	989 (78.37)	730 (90.68)
≥ **6.11**	296 (3.44)	748 (25.48)	273 (21.63)	75 (9.32)
Fructosamine (mmol/L)				
Mean(SD)	2.04 (0.18)	2.22 (0.37)	2.15 (0.34)	2.05 (0.28)
< 2.6	8597 (99.83)	2619 (89.20)	1191 (94.37)	777 (96.52)
≥ **2.6**	15 (0.17)	317 (10.80)	71 (5.63)	28 (3.48)
Total Cholesterol (mmol/L)				
Mean(SD)	5.62 (1.10)	6.43 (1.13)	5.99 (1.26)	5.42 (1.14)
< 6.50	6753 (78.41)	1530 (52.11)	837 (66.32)	654 (81.24)
≥ 6.50	1859 (21.59)	1406 (47.89)	425 (33.68)	151 (18.76)
HDL Cholesterol (mmol/L)				
Mean(SD)	1.69 (0.36)	1.18 (0.39)	1.32 (0.42)	1.50 (0.40)
< 1.03	144 (1.67)	950 (32.36)	272 (21.55)	91 (11.30)
≥ 1.03	8468 (98.33)	1986 (67.64)	990 (78.45)	714 (88.70)
LDL Cholesterol (mmol/L)				
Mean(SD)	3.45 (0.98)	4.21 (1.06)	3.79 (1.09)	3.38 (1.05)
< 4.10	6551(76.07)	1362 (46.39)	805 (63.79)	627 (77.89)

≥ **4.10**	2061 (23.93)	1574 (53.61)	457 (36.21)	178 (22.11)
Triglycerides (mmol/L)				
Mean(SD)	1.06 (0.47)	2.35 (1.25)	1.99 (1.32)	1.33 (0.84)
< 1.71	7843 (91.07)	993 (33.82)	653 (51.74)	639 (79.38)
≥ 1.71	769 (8.93)	1943 (66.18)	609 (48.26)	166 (20.62)
Apolipoprotein A-1 (g/L)				
Mean(SD)	1.48 (0.22)	1.33 (0.20)	1.38 (0.22)	1.40 (0.22)
< 1.05	88 (1.02)	157 (5.35)	52 (4.12)	31 (3.85)
≥ 1.05	8524 (98.98)	2779 (94.65)	1210 (95.88)	774 (96.15)
Apolipoprotein B (g/L)				
Mean(SD)	1.11 (0.29)	1.52 (0.34)	1.34 (0.38)	1.14 (0.33)
< 1.50	7828 (90.90)	1499 (51.06)	875 (69.33)	700 (86.96)
≥ **1.50**	784 (9.10)	1437 (48.94)	387 (30.67)	105 (13.04)
Log (triglycerides/HDL) c				
mean(SD)	(-)0.54 (0.53)	0.63 (0.79)	0.29 (0.87)	(-)0.23 (0.75)
< 0.5	8576 (99.58)	1166 (39.71)	770 (61.01)	685 (85.09)
≥ 0.5	36 (0.42)	1770 (60.29)	492 (38.99)	120 (14.91)
ApoB/ApoA-I c				
mean(SD)	0.76 (0.22)	1.15 (0.28)	0.99 (0.31)	0.83 (0.28)
< 1.00	7582 (88.04)	699 (23.81)	683 (54.12)	620 (77.02)
≥ 1.00	1030 (11.96)	2237 (76.19)	579 (45.88)	185 (22.98)
AST (IU/L)				
Mean(SD)	20.05 (6.55)	21.24 (6.49)	46.41 (52.16)	21.61 (20.46)
< 45	8565 (99.45)	2925 (99.63)	876 (69.41)	789 (98.01)
≥ 45	47 (0.55)	11 (0.37)	386 (30.59)	16 (1.99)
ALT (IU/L)				
Mean(SD)	21.60 (9.35)	28.32 (10.02)	84.71 (91.11)	23.59 (12.90)
< 50	8578 (99.61)	2936 (100.00)	0 (0.00)	782 (97.14)
≥ **50**	34 (0.39)	0 (0.00)	1262 (100.00)	23 (0.17)
GGT (IU/L)				
Mean(SD)	22.93 (23.71)	38.18 (31.99)	93.70 (117.85)	30.24 (26.33)
Normal (<18)	4713 (54.73)	447 (15.22)	33 (2.61)	318 (39.50)
Normal high (18-36)	2959 (34.36)	1453 (49.49)	284 (22.50)	287 (35.65)
Elevated (36-72)	731 (8.49)	780 (26.57)	442 (35.02)	145 (18.01)
Highly elevated (>72)	209 (2.43)	256 (8.72)	503 (39.86)	55 (6.83)
Creatinine (µmol/L)				
Mean(SD)	77.42 (15.03)	84.80 (17.99)	83.17 (13.92)	79.22 (18.58)
Low	20 (0.23)	11 (0.37)	5 (0.40)	4 (0.50)
Normal	7863 (91.30)	2412 (82.15)	1108 (87.80)	705 (87.58)
High	729 (8.46)	513 (17.47)	149 (11.81)	96 (11.93)
Albumin (g/L)				
Mean(SD)	42.98 (2.71)	43.14 (2.72)	44.09 (3.01)	41.79 (3.31)
<35	6 (0.07)	7 (0.24)	3 (0.24)	12 (1.49)
>35	8606 (99.93)	2929 (99.76)	1259 (99.76)	793 (98.51)
Leukocytes (10^9 cells/L)				
Mean(SD)	6.24 (1.79)	6.89 (1.99)	6.55 (2.14)	7.56 (2.64)
<10	8374 (97.24)	2703 (92.06)	1206 (95.56)	673 (83.60)
≥ **10**	238 (2.76)	233 (7.94)	56 (4.44)	132 (16.40)
C-Reactive Protein (mg/L)				
Mean(SD)	4.99 (15.41)	5.51 (12.17)	5.78 (7.53)	16.63 (24.12)
<10	7764 (90.15)	2557 (87.09)	1095 (86.77)	442 (54.91)
10 15	717 (8.33)	265 (9.03)	106 (8.40)	108 (13.42)
15-25	62 (0.72)	78 (2.66)	33 (2.61)	92 (11.43)

25-50	48 (0.56)	31 (1.06)	23 (1.82)	98 (12.17)
>50	21 (0.24)	5 (0.17)	5 (0.40)	65 (8.07)
Iron (µmol/L)				
Mean(SD)	18.77 (5.43)	17.99 (4.70)	20.00 (6.65)	8.86 (2.79)
Low	1 (0.01)	4 (0.14)	36 (2.85)	595 (73.91)
Normal	8276 (96.10)	2888 (98.37)	1138 (90.17)	210 (26.09)
High	335 (3.89)	44 (1.50)	88 (6.97)	0 (0.00)
TIBC (mg/dL)				
Mean(SD)	0.32 (0.10)	0.30 (0.09)	0.33 (0.12)	0.15 (0.05)
Low	2050 (23.80)	877 (29.87)	335 (26.55)	805 (100.00)
Normal	4490 (52.14)	1603 (54.60)	557 (44.14)	0 (0.00)
High	2072 (24.06)	456 (15.53)	370 (29.32)	0 (0.00)
Phosphate (mmol/L)				
Mean(SD)	1.07 (0.16)	1.06 (0.18)	1.07 (0.19)	1.10 (0.19)
Low	70 (0.81)	7 (0.24)	8 (0.63)	10 (1.24)
Normal	8160 (94.75)	2750 (93.66)	1185 (93.90)	701 (87.08)
High	382 (4.44)	179 (6.10)	69 (5.47)	94 (11.68)
Calcium (mmol/L)				
Mean(SD)	2.37 (0.09)	2.39 (0.10)	2.40 (0.10)	2.35 (0.10)
Low	110 (1.28)	26 (0.89)	13 (1.03)	42 (5.22)
Normal	8394 (97.47)	2838 (96.66)	1212 (96.04)	751 (93.29)
High	108 (1.25)	72 (2.45)	37 (2.93)	12 (1.49)
Life Status				
Alive	6815 (79.13)	2066 (70.37)	973 (77.10)	603 (74.91)
Death	1797 (20.87)	870 (29.63)	289 (22.90)	202 (25.09)
Cancer	1148 (13.33)	490 (16.69)	197 (15.61)	121 (15.03)

The following abbreviations have been used in Table 1: High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Apolipoprotein A-1 (ApoA), Apolipoprotein B (ApoB), Gamma-Glutamyl transferase (GGT), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST) and Total iron binding capacity (TIBC).

+The missing values are not included in the percentage of the Educational Status categories.

Table 20. MODDEL H. Hazard ratios and 95 % confidence interval for the association of LCA-derived metabolic classes and overall cancer risk and cancer specific risk.

	Hazard Ratios (95% CI) [a]	Hazard Ratios (95% CI) [b]
Cancer Risk: All cancer types		
Number of events	1,956	1,956
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.09 (0.98 - 1.22)	1.05 (0.94 - 1.17)
3 - Liver	1.28 (1.10 - 1.50)	1.28 (1.09 - 1.49)
4 – Inflammation & Iron	1.17 (0.97 - 1.41)	1.17 (0.97 - 1.41)
Cancer Risk: Buccal cavity and pharynx		
Number of events	34	34
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.79 (0.77 - 4.14)	1.70 (0.73 - 1.17)
3 - Liver	2.66 (0.96 - 7.35)	2.60 (0.94 - 7.16)
4 - Inflammation & Iron	3.94 (1.38 - 11.30)	3.77 (1.31 - 10.82)
Cancer Risk: Digestive organs and peritoneum		
Number of events	330	330
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	0.83 (0.62 - 1.11)	0.83 (0.62 - 1.11)
3 - Liver	2.12 (1.54 - 2.91)	2.12 (1.54 - 2.91)
4 - Inflammation & Iron	0.86 (0.51 - 1.46)	0.86 (0.51 - 1.46)
Cancer Risk: Respiratory system		
Number of events	133	133
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.40 (0.94 - 2.08)	1.32 (0.88 -1.96)
3 - Liver	0.90 (0.44 - 1.82)	0.87 (0.43 - 1.77)
4 - Inflammation & Iron	1.48 (0.76 - 2.88)	1.46 (0.75 - 2.84)
Cancer Risk: Skin melanoma		
Number of events	205	205
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	0.78 (0.56 - 1.10)	0.78 (0.56 - 1.11)
3 - Liver	0.71 (0.40 - 1.26)	0.73 (0.41 - 1.31)
4 - Inflammation & Iron	0.70 (0.35 - 1.37)	0.70 (0.35 - 1.37)
Cancer Risk: Breast and genito-urinary organs		
Number of events	655	655
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.19 (0.99 - 1.42)	1.12 (0.94 - 1.33)
3 - Liver	1.04 (0.80 - 1.37)	1.04 (0.80 - 1.37)
4 - Inflammation & Iron	1.25 (0.91 - 1.71)	1.25 (0.91 - 1.71)

Cancer Risk: Brain & nervous system, Thyroids		
Number of events	34	34
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.01 (0.51 - 1.99)	0.96 (0.48 - 1.00)
3 - Liver	1.01 (0.38 - 2.67)	0.99 (0.38 - 2.59)
4 - Inflammation & Iron	0.92 (0.28 - 2.99)	0.91 (0.28 - 2.96)
Cancer Risk: Connective and endocrine tissue		
Number of events	56	56
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	0.65 (0.21 - 1.95)	0.64 (0.21 - 1.94)
3 - Liver	2.65 (1.00 - 7.02)	2.67 (1.01 - 7.07)
4 - Inflammation & Iron	3.00 (1.11 - 8.11)	2.96 (1.10 - 8.00)
Cancer Risk: Lymphatic and hematopoietic tissues: Hodgkin lymphoma, Non-H lymphoma, Leukaemia and Myeloma		
Number of events	129	129
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.72 (1.15 - 2.56)	1.68 (1.12 - 2.51)
3 - Liver	1.65 (0.91 - 3.00)	1.68 (0.93 - 3.05)
4 - Inflammation & Iron	1.23 (0.56 - 2.68)	1.25 (0.57 - 2.73)

a Time scale adjusted for age, sex and CCI
b Age scale adjusted for sex and CCI

Table 21. MODEL H. Hazard ratios and 95 % confidence interval for the association of LCA- derived metabolic classes and all causes death and cancer death.

	Hazard Ratios (95% CI) [a]	**Hazard Ratios (95% CI) [b]**
All causes death		
Number of events	3158	3158
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.26 (1.16 - 1.37)	1.29 (1.19 - 1.40)
3 - Liver	1.67 (1.47 - 1.90)	1.70 (1.49 - 1.93)
4 - Inflammation & Iron	1.21 (1.05 - 1.41)	1.20 (1.04 - 1.40)
Cancer death		
Number of events	706	706
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.22 (1.02 - 1.45)	1.20 (1.01 - 1.42)
3 - Liver	1.44 (1.11 - 1.86)	1.46 (1.13 - 1.90)
4 - Inflammation & Iron	0.93 (0.66 - 1.32)	0.93 (0.66 - 1.32)

a Time scale adjusted for age, sex and CCI
b Age scale adjusted for sex and CCI

Table 22. C statistics for the MODEL H. LCA metabolic profiles and the standard health markers Total Cholesterol, Glucose and Gamma Glutamyl Transferase were assessed for all the outcomes studied: cancer, cancer death and overall death using calendar-time as a time scale.

	Cancer Risk		**Cancer Death Risk**		**Overall Death Risk**	
	HR (95% CI) adjusted for Age,Sex and CCI	**c statistics**	**HR (95% CI) adjusted for Age,Sex and CCI**	**c statistics**	**HR (95% CI) adjusted for Age,Sex and CCI**	**c statistics**
MODEL H:						
1- Normal Class (60%)	1.00 (ref)	0.70	1.00 (ref)	0.77	1.00 (ref)	0.84
2- Lipids (23%)	1.09 (0.98 - 1.22)		1.22 (1.02 - 1.45)		1.26 (1.16 - 1.37)	
3- Liver (9%)	1.28 (1.10 - 1.50)		1.44 (1.11 - 1.86)		1.67 (1.47 - 1.90)	
4 - Iron (8%)	1.17 (0.97 - 1.41)		0.93 (0.66 - 1.32)		1.21 (1.05 - 1.41)	
Single Biomarkers:						
Total Cholesterol	1.00 (0.96 - 1.05)	0.69	1.03 (0.96 - 1.10)	0.76	0.99 (0.96 -1.02)	0.83
Glucose	1.01 (0.87 - 1.16)	0.69	1.03 (0.83 - 1.29)	0.76	1.38 (1.26 - 1.52)	0.83
Gamma Glutamyl Transferase	1.06 (1.02 - 1.11)	0.69	1.11 (1.05 - 1.17)	0.76	1.14 (1.11 - 1.17)	0.83

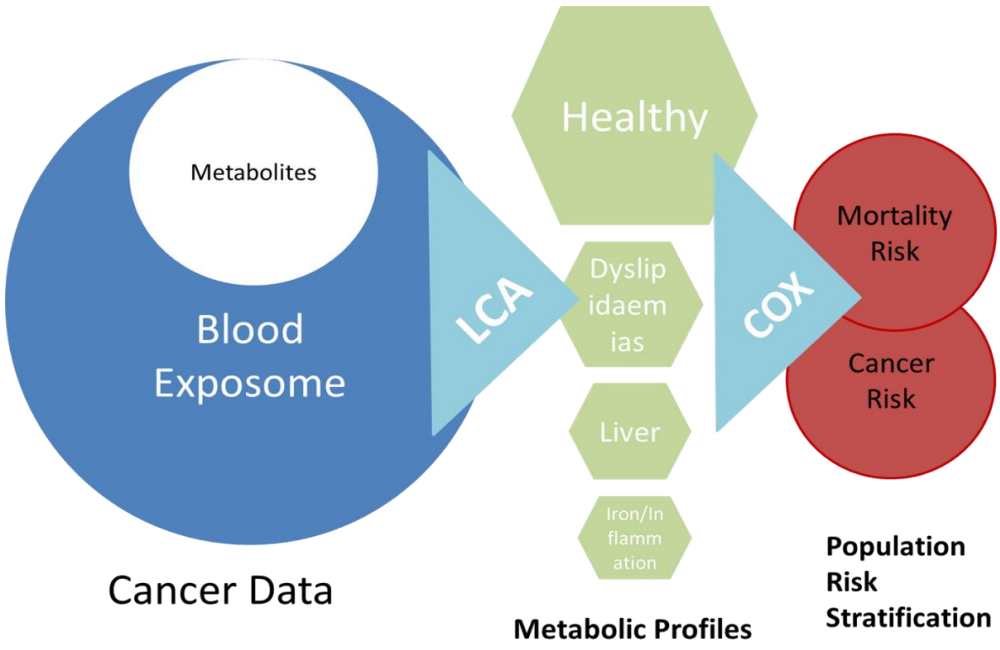
d. Discussion

To investigate population heterogeneity and cancer susceptibility in a given population, I aimed to identify novel statistical approaches to improve stratification of individuals, based on their underlying risk of developing cancer and risk of increasing mortality, by exploring the stratification capabilities of multiple markers of the blood exposome. The optimised multistage methodology employed in this project, indicated that standard of care baseline serum markers when assembled into meaningful metabolic profiles by data reduction techniques, can help stratify the population for cancer risk, cancer mortality and overall mortality. More specifically, I observed that abnormal values for markers of the lipid metabolism, liver function and inflammatory and iron metabolism distinguish participants into metabolic profiles, which were associated with of long term cancer risk and/or mortality and cancer mortality. Moreover, the results suggested that data reduction methods, specifically latent class analysis, can slightly improve the stratification models for cancer and mortality in comparison with single standard health biomarkers (e.g.: total cholesterol, glucose and GGT) in biomedical settings via

extraction of the population heterogeneity hidden in the biological data. Even though the results are quite comparable, to assess improvement in the predictions a formal testing of the statistical significance would have been required.

The outcome of the statistical pipeline that explored the blood exposome to assess population heterogeneity associated with cancer risk and mortality is illustrated in Figure 42.

Figure 42. Study statistical pipeline describing the methodology followed in the project and LCA outcome.



Metabolic profiles

Among the biological pathways addressed in the LCA for the final model H, abnormalities in the lipid metabolism were the most common. Hyperlipidaemia was present in about a quarter of the study population explaining the largest abnormal metabolic profile. The weight of the lipid profile in the analysis was consistent with the reported global prevalence of hypercholesterolemia among adults (37% for males and 40% for females) as reported in the Global Health Observatory in 2008 estimates by the World Health Organization (WHO) and the results from the

Swedish population in the WHO MONICA project (286). Dyslipidaemia is associated with higher risk of CVD and other chronic diseases such as cancer, as also observed in the study (287). Liver dysfunction, iron deficiency and altered inflammatory markers profiles also distinguished important subgroups in the study population. About 9% of the population had abnormal values for markers of liver functioning (GGT, AST and ALT), which is similar to the results obtained in a population-based survey in the United States that estimated abnormal alanine aminotransferase (ALT) was present in 9% of respondents in absence of viral hepatitis C or excessive alcohol consumption (288). Moreover, these enzymes are known to be linked to cancer because of their role in preserving the intracellular homeostasis of the oxidative stress (201, 289, 290), which is concordant with the results of these analyses. The iron profile and inflammatory markers clustered 6% of individuals in the study, which was predominantly driven by low levels of serum iron and TIBC, as well as high levels of CRP and leukocytes. This could potentially point towards anaemia of inflammation, a chronic inflammation presenting low iron values, that occurs because the iron deficiency provides the body with infection resistance, which demonstrates the tightly connection between the inflammatory response and the iron and its homeostasis (291). This condition has been reported in more than 30% of cancer patients at time of diagnosis (292-294).

Metabolic profiles as a risk factor for long term cancer and mortality

The above-described three classes of abnormal metabolic profiles were all associated with an increased risk of cancer and worse survival, as compared to the healthy class. The findings therefore confirm the key importance of these metabolisms and the specific molecules in the maintenance of the intracellular homeostasis and how their unbalance can be related with the aetiology of cancer disease and mortality (3). The LCA adapted in this study thus illustrates how a biomarker-wide approach can help assess markers of the blood exposome in the context of carcinogenesis and mortality (117) (Figure 41).

More specifically, individuals presenting **abnormal liver function** markers carried worse outcomes in terms of overall cancer risk and cancer death, and a positive association with digestive, connective and endocrine cancers diagnosis. Moreover, the participants with this profile had a higher probability of overall death. These results are consistent with previous published data. A positive association between elevated GGT and overall cancer risk, with no interaction of ALT, was found in the AMORIS cohort previously (172), and it was also reported in other large cohort studies (295, 296). These studies also found strong associations with elevated levels of GGT and digestive and respiratory cancer incidence. Elevated GGT has been associated with mortality from all causes, liver disease, cancer and diabetes, while ALT only showed associations with liver disease death in a large US cohort (297). However, in a study based on an elderly population it was found that GGT was associated with increased cardiovascular disease mortality, and ALP and AST with increased cancer-related mortality (298). Moreover, a meta-analysis evaluating the associations between liver enzymes and all-cause mortality found positive independent associations of baseline levels of GGT and ALP with all-cause mortality (299). Furthermore, some of the multiple functions of the liver organ include to process the nutrients absorbed from the small intestine in the digestive system, to produce important hormones of the endocrine system such as Insulin-Like Growth Factor 1 and an active role in the maintenance of tolerance to self-molecules which pointed it as target for autoimmune diseases as connective tissue diseases (CTD) (300), therefore underlying biological links exists between the altered abnormal function and the association with the digestive, connective and endocrine cancers.

In the present study, the liver biomarker profile was positive associated with all the outcomes studied, suggesting a key role of this pathway in the development of cancer, probably related with its active role maintaining the intracellular redox regulation. Moreover, the specific enzymes studied in the project participate in relevant body functions, that according to the hallmarks of cancer publication are fundamental features in the development of cancer, such as gluconeogenesis cycle (ALT), amino acid metabolism (AST) and xenobiotic detoxification pathway (GGT) (4,

29). Thus, further investigations are necessary to establish the potential of the altered enzymes and the liver profile as a tool for cancer risk stratification.

Individuals allocated to the **lipid profile** presented positive associations with cancer mortality, and overall mortality and higher risk of lymphatic and hematopoietic cancers. The link between hyperlipidaemia and mortality has been studied broadly, with associations with established links for cancer and all-cause mortality (301-303). The association between lipids and lymphatic and hematopoietic cancers is more controversial, as some studies found an inverse association for these cancers and high levels of serum cholesterol (304, 305) and a systematic literature review from 2016 found no association (306). However, the fatty acid metabolism and hematopoietic cell biology are connected through the WNT signalling pathway, an established dysregulated pathway in cancer(22), and its association has been investigated for leukaemia (307). Moreover, recently has been suggested the active and intricated lymphatic role in lipid transportation (308).

Participants clustered in the **unbalanced iron profile and inflammation** had an increased risk of endocrine, buccal and oral cancers and were observed to have a higher risk of all-causes death. Altered inflammation and iron metabolisms are key metabolic 'hallmarks of cancer' (3, 179, 309). These markers are tightly connected via a body response denominated as anaemia of inflammation that is characterised for body iron sequestration in response to an underlying inflammatory disease or malignancy (294, 310). Moreover, CRP and leukocytes have been associated with cancer risk in AMORIS previously (170) and the observation of an association with an increased risk of buccal and oral cancer corroborates previous findings in AMORIS (179). Furthermore, a study in oral cancer patients suggested that iron deficiency is associated with increased oxidative stress, increasing the risk of oral cavity cancer (311). In relation with all-causes death, iron deficiency, an indicator of malnutrition considered to contribute to maternal, perinatal and over-all death, ranked 9 out of 22 mortality risk on to the Global burden of Disease 2000 publication, accounting for 841,000 deaths per year (312, 313).

Population heterogeneity and risk stratification: the need for data reduction techniques

The modulation effect of population heterogeneity on the association between potential risks factors and disease is a new avenue to understand the variability of risk in the population (314). Moreover, given the multifactorial nature of chronic diseases influenced by diverse exposures and the potential of the blood exposome to unravel cancer susceptibility in a population, a new area for innovative methodological approaches investigating disease risk stratification to improve disease prevention, has emerged. For instance, in a targeted metabolomics exercise Shan et al. performed a principal component analysis and time to event analysis identifying metabolic profiles to predict risk of CVD (275). Another study used Monte Carlo Cross Validation and Lasso logistic regression to evaluate serum biomarkers as an alternative to faecal immunochemical testing to improve detection of colorectal cancer (118). In 2010, the European Prospective Investigation on Cancer and Nutrition (EPIC) cohort reported that a specific prediagnostic plasma phospholipid fatty acid profile could predict the risk of gastric cancer (270). As rationalized in the HELIX project, these multiple profiling approaches aim to identify groups of individuals in the population that share a similar exposome signatures that might account for differences on the specific risk of study (111). Together with these studies, the systematic data integration approach based on LCA presents the potential of investigating population heterogeneity using metabolic profiling, as a subset of the exposome, as risk factors for long term cancer risk and mortality. The LCA metabolic profiles can be further explored as markers of cancer and mortality susceptibility, however their prediction capabilities were very comparable with established health markers such as total cholesterol, glucose or gamma glutamyl transferase.

Future steps

This project, as a prospective cohort study, examined the association of the markers with cancer disease and mortality. The biomarkers investigated in the project have been explored as markers of susceptibility, given that the measurements included were taken three years before any outcome event, so these pre-clinical biological samples are not influenced by the inherent metabolic changes due to the disease itself. In principle, the measurements would not be affected by disease at baseline for most of cancer types, given the cancer lack time, however these markers would reflect both the exposure and early biological effects of disease (Figure 8).

Markers of susceptibility are valuable and effective instruments for cancer stratification on screening and prevention programmes, however in order to establish their prediction capabilities (Table 22) further studies to validate the results whilst allowing to measure sensitivity and specificity, will need to be conducted. Some examples of such studies are the pilot studies conducted in the EPIC cohort in relation to colon, breast and hepatocellular cancer outcome (244, 272). Hence, a nested case-control could be conducted in AMORIS to determine the predictive capabilities of the metabolic profiles and markers that defined the profiles to estimate cancer risk and mortality. Moreover, a validation exercise in a different study population characterised by diverse life-style, demographics and genetics, such as the American National Health and Nutrition Examination Survey (NHANES), would allow to establish the external validity of the metabolic profiles without compromising the validity of the statistical approach (288, 315-318). To establish the external validity of the LCA metabolic derived classes, a validation exercise has already been performed using data from NHANES based on the same panel of serum markers – with additional adjustments for alcohol and smoking consumption. The results showed the same four LCA metabolic derived classes as seen in AMORIS: one normal class with normal values for all the markers (66% population), one class defined by abnormal values for the lipids ratios (22%), one class strongly defined for abnormal values of inflammation markers and low iron (9%) and a small class defined by abnormal values of liver enzymes (3%). The

NHANES database lacks information on date of cancer diagnosis and so therefore these analyses were limited to overall death and cancer-specific death as an outcome (*Note: this work was conducted outside the scope of the PhD; manuscript in preparation)*. In this analysis the iron and inflammation class were strongly associated with both outcomes (Overall death HR (95% CI) =1.99 (1.09 - 3.62) and Cancer Death HR (95% CI) = 2.93 (1.14 - 7.54)), as compared to the normal class. The rest of the classes were not associated with the outcomes. The NHANES findings thus implies general validity of the methodology used to explore population heterogeneity in the context of metabolic profiles and cancer susceptibility.

Future studies will benefit from longitudinal approaches assessing repeated multiple serum marker measurements that will capture the population phenotypic variations in relation to disease over long periods of time and will help understand the complexity of cancer and be able to stratify population at risk of cancer and mortality.

Strengths and limitations

The present study was conducted in a large and well-defined population, applying a multi-faced approach covering main biological pathways to assess biomarker profiles that could indicate cancer risk, cancer survival and mortality. The major strength of these analyses lies in the innovative avenue to study population heterogeneity and susceptibility to disease and mortality in a large cohort of participants with multiple routinely collected blood biomarkers (a sub-set of the exposome), all measured on fresh blood samples on the same day at the same clinical laboratory. I included all the markers available in the cohort for a large population (n>13000) in an exploratory data driven approach, however not every marker of the central metabolic pathways was available in the database (i.e. Complete Blood Count). Please see project A for a further discussion about the generalisability of the AMORIS cohort. The main limitation of the study is the partial availability of information on life-style and environmental factors established as cancer risk factors such as BMI, smoking, alcohol, diet and hypertension. The lack of

these external exposures compromised the extent of the markers utilised to characterize the external exposome component and limited the characterisation of the metabolic profiles in the population. However, to mitigate the lack of external factors, the analyses have been adjusted for Charlson Comorbidity Index which includes comorbidities such as obesity and hypertension and other chronic diseases such as circulatory or respiratory diseases. Alcohol consumption is also an important risk factor for cancer; the lack of information on these variables was mitigated by using information on serum biomarkers such as gamma glutamyl transferase and other liver enzymes (319, 320). The rather low number of respiratory cancers in this population may be explained by a lower prevalence of smoking in Sweden at the time (15). Moreover, given the results of the validation study in NHANES, which did include adjustment for both alcohol and smoking consumption, it can be suggested that lack of adjustment for these confounders has little effect on the actual association between the latent classes and the outcomes studied.

Conclusion

The LCA adapted in this study illustrates how an exploratory biomarker-wide approach, can help to assess population susceptibility to disease and provide insight into disease aetiology in the context of carcinogenesis and mortality. The analyses outlined the importance of the lipid molecules, the enzymes ALT, AST and GGT and the iron metabolites when classifying individuals to evaluate hidden heterogeneity based on blood metabolites. These findings suggest the relevant function of these established biomarkers in the internal body homeostasis, however these molecules' utility is commonly reduced to the specific biological pathways linked to their current clinical application (i.e.: standard health assessment). Considering the tight interaction of the internal metabolic molecules and the external exposome and given the environmental and genetic modulation of metabolic molecules, metabolic profiling based on standard of care serum markers could become a useful non-invasive predictive signature for risk stratification and an important area of research for mechanisms and clinical relevance. Further studies, including larger panels of

markers and life style and environmental factors will be necessary to establish the potential role of these methods in stratified medicine.

Therefore, the findings support the recently expressed need for a shift from the classical epidemiological approach of assessing one exposure to a systemic approach with multiple exposures in order to capture the population and disease heterogeneity and better characterise the population cancer susceptibility with the ultimate aim of improving prevention and early detection of cancer.

Chapter V. Individual susceptibility to Disease.

The findings of this section were published as an abstract in the European Journal of Surgical Oncology (EJSO) November 2015 (Appendix III) and published in Biomedical Optical Express June 2017 (321) (Appendix IV).

To accomplish personalised medicine in cancer, acquiring a comprehensive perspective of the population susceptibility to cancer, in conjunction with insight knowledge of the inherent variability of the disease, will be necessary.

In the previous chapter, population heterogeneity was investigated in a subset of blood exposome markers which when reduced to meaningful metabolic components, using data reduction techniques, presented the capability to stratify the population for cancer risk and mortality. This confirms that efforts towards exploring population cancer susceptibility could help us to optimise stratification models for prevention and early detection of the disease.

Therefore, this chapter investigates the heterogeneity inherent to cancer in a translational imaging setting. Focusing on breast cancer, the project aimed to distinguish variability at cellular level by differentiating tissue composition (tumour, fibrous and adipose material) using imaging data generated in an ex vivo study (REC 12-EE-0493). Considering the complexity and multidimensionality of imaging data, the data reduction techniques were applied to perform supervised clustering allocating each observation to a specific tissue type through a predictive analysis.

To investigate individual susceptibility to cancer in an imaging setting, I explored individuals' heterogeneity in cancer composition in a diagnostic context to improve breast cancer tissue discrimination intraoperatively, using material from King's Health Partners' Breast Cancer Biobank:

Project C: I evaluated how imaging data produced by a terahertz probe can distinguish between different breast tissue samples, with the ultimate aim of using this probe intraoperatively in breast-conserving surgery to predict positive tumour margins.

I. Project C: Discrimination of breast tumour tissue

a. Rationale

Breast cancer (BC) is the leading cause of cancer and cancer death in females worldwide, accounting for about 2.4 million cancer BC cases and about 523,000 BC deaths in 2015 (15) (Figure 3).

Surgery to remove the primary tumour is one of the main treatment options for BC patients, which has evolved from radical surgery options (mastectomy) to more conservative approaches, such as breast-conserving surgery (BCS) in the last decades. Since 1990, BCS has become a patient's treatment of choice, because of the cosmetic and less invasive implications, estimating that in United Kingdom two-thirds of newly diagnosed patients undergo BCS as initial treatment (322). Currently, BCS is also performed in large or locally advanced tumour cases after pre-surgery tumour downstaging and downsizing using neoadjuvant chemotherapy or endocrine therapy treatments (323-325).

To determine the surgical outcome of BCS procedures, the assessment of the negative tumour margins in the excised material post-surgery, is a common practice (324). Moreover, about 20% of the patients that undergo a BCS require a re-operation due to the positive tumour margins on a postoperative pathological assessment, which is considered the gold standard for estimating tumour margins (326). The additional surgery to obtain clear margins and/or remove remaining lymph nodes in the axilla, causes a significant physical and psychological morbidity on patients and an economic burden in health systems (327-329).

Hence, in an effort to decrease re-excision rates, different reliable intraoperative margin assessment (IMA) tools are being investigated to assess tumour resection margins intraoperatively. Some of these techniques are clinically established, such as frozen section analysis (330, 331), specimen radiography (332), intraoperative ultrasound (333) and touch imprint cytology (331), however these IMAs present diverse performance and limitations in terms of accuracy, speed, cost, and reliability. Therefore, new IMA tools are currently under development including Raman spectroscopy (334), microcomputed CT (335), mass spectroscopy (336) or fluorescence imaging (337, 338).

Terahertz pulsed imaging (TPI) is such a developing IMA tool, that employs terahertz (THz) radiation with very low proton energy (0.1 – 4 THz) which does not cause harmful ionization in biological tissues (339, 340). Given its the millimetric penetration depth and sensitivity of THz radiation to changes in water content and tissue composition, it has been applied to image biological tissue in different studies, such as in cancer research to distinguish malignant from benign tissue in skin, colon, oral, gastric, brain and breast cancer in 2006 (341, 342).

As a consequence of the previous studies, Teraview Ltd. (Cambridge, UK) developed a handheld THz Probe to intraoperatively assess tumour margins and sentinel lymph nodes in breast cancer patients, eventually aiming to reduce the number of BCS reoperations. A collaborative feasibility study was established between Teraview Ltd. (Cambridge, UK) and Prof Purushotham's research team at King's Health Partners (KHP) to test the performance of the probe in ex-vivo fresh breast tissue (REC 12-EE-0493).

Thus, to understand each individuals' susceptibility to cancer driven by disease variability, whilst evaluating the application of data reduction methods in an imaging setting, I used data generated in this collaborative pilot study of Teraview and KHP. The project aimed to distinguish between different breast tissue samples, based on the imaging data produced by the probe, to ultimately identify malignant

tissue intraoperatively ensuring clear negative tumour margins in BCS. Therefore, a supervised clustering method, the Naïve Bayes Classifier algorithm was utilised to build a classifier able to distinguish between tumour, fibrous and adipose tissue in real – time surgery.

b. Methods

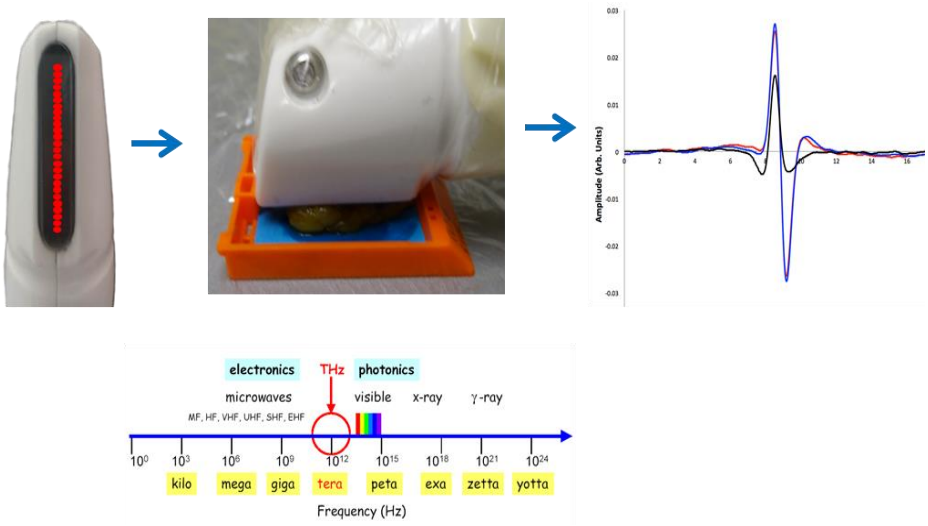
i. Study population: Data collection

The data included in the project was obtained from the first-in-human, single centre pilot study to evaluate the ability of a handheld TPI probe to discriminate benign from malignant breast tissue in an ex vivo setting conducted in KHP.

Technology

The study utilised a terahertz pulsed imaging (TPI) handheld probe device version 2.0 (Teraview Ltd. Cambridge, UK). The TPI probe produced THZ pulses of frequency 0.1-2.0 THz through a scan window that contained 26 pixels, providing a pixel resolution of about 0.6 mm. During the tissue scanning, each pixel received residual THZ pulses from the tissue that formed TPI waveforms as illustrated in Figure 43.

Figure 43. Schematic description of the raw data acquisition. Images are courtesy of the study (REC 12-EE-0493). *TPI handheld probe measurement of tissue sample positioned in histology cassette. Residual THZ pulses are received by each pixel from the tissue producing typical TPI waveforms per pixel. Based on the type of tissue present in the breast sample the TPI waveform presented a different shape (tumour (blue), fibrous (red), and adipose cells (black)) (image on the right).*



Data acquisition

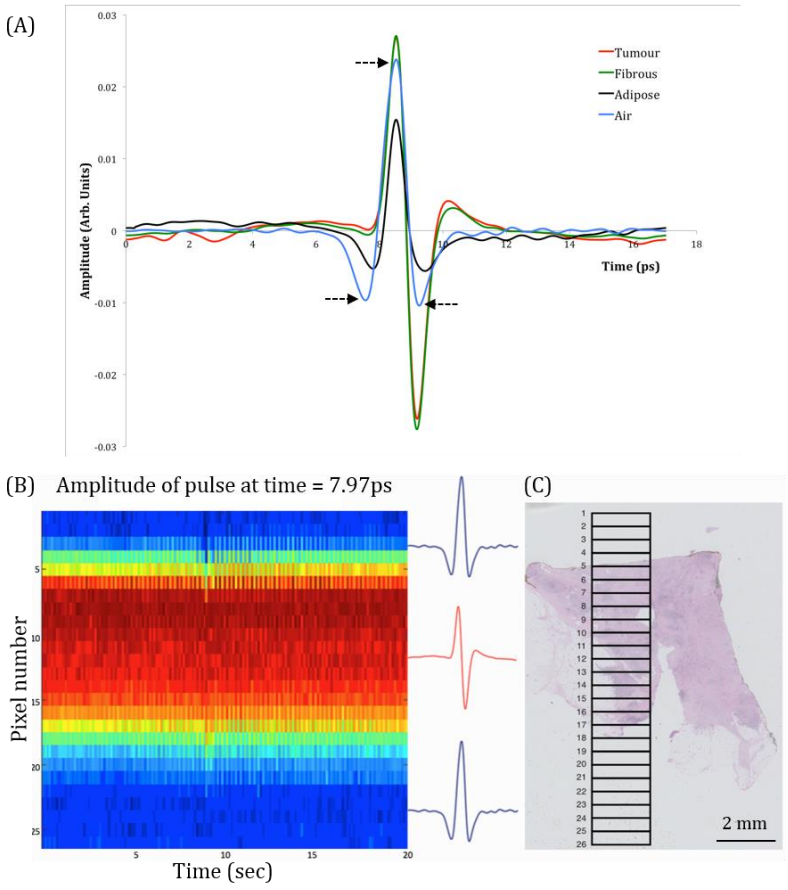
Imaging data was generated by scanning, with the TPI handheld probe, 46 freshly excised breast cancer samples from 30 BC patients, who underwent BCS or mastectomy at Guy's Hospital (Guys and St Thomas NHS Foundation Trust) between August 2013 and August 2014. Written informed consent was obtained for the research use of the material (REC 12-EE-0493).

The data acquisition followed the following pathway: (i) while the ex vivo breast tissue was scanned with the probe, using air as a reference, (ii) each pixel transmitted many THZ pulses to the tissue sample and (iii) each pixel received multiple residual THZs per scan of tissue sample, following (iv) the residual THZs were averaged into a TPI waveform with 674 data points of information per pixel within the device software, resulting in (v) the final raw data obtained from the device that contained one waveform with 674 data points per pixel producing a matrix of 26 waveforms (one per pixel) with 674 data points of information per scan of a tissue sample. A posterior pre-processing of the data to remove the noise of the signals, reduced the number of data points to 301 per waveform (final matrix of n=26 waveforms x m=301 data points per scan of tissue sample).

Moreover, detailed histopathology was correlated with the Terahertz time domain data (THz) per pixel area by two pathologists blinded for the patient's details. Consequently, each waveform per pixel was labelled with the percentage of tumour, fibrous and adipose content (Figure 44). The ex vivo breast tissue samples were characterised based on their tissue composition which included different proportions of tumour tissue, consisting of tumour cells embedded with normal breast cells, adipose tissue, mainly composed by adipocytes cells and fibrous or connective tissue containing elastic and collagenous fibres, ground substance and cells. The 46 tissue samples included in the project had the following tissue composition: (i) 20 samples contained tumour including 16 invasive ductal (IDC)/no special type (NST) carcinoma, two NST admixed with ductal carcinoma in situ (DCIS), four invasive lobular carcinoma (ILC), (ii) 22 samples contained pure fibrous tissue

or a mixture of fibrous and adipose tissue (78% characterised by its high-fibrous density content), and (iii) four samples contained pure adipose tissue.

Figure 44. Correlating TPI waveforms with histopathology. (A) Typical impulse function of tumour, fibrous, and adipose tissue, and air, respectively. Clear differences are seen between the impulse functions from air and from tissue, and between adipose and tumour/fibrous tissue (black arrows). (B) TPI image from sample based on the amplitude of the impulse function at t = 7.97ps. (C) Digital histopathology slide of the same tissue sample. By using the photograph of the sample in combination with the air-tissue interface visible in the TPI image, the TPI 15 x 2 mm scan area can be accurately mapped onto the histopathology slide (black rectangle). The pixels are displayed as intermittent horizontal lines at 0.6 mm distance in the scan window. Pixel 5 – 17 contain invasive ductal/no special type (NST) carcinoma; the percentage of tumour cells in each pixel area ranges between 5 – 10%. The tissue immediately surrounding the tumour cells (called background) is composed of fibrous tissue, whilst fatty adipose tissue is seen inferiorly. Figure and caption taken from *Grootendorst et al. 2017* (321).



Final dataset

The final dataset contained selected THz data from each tissue sample, given that the size and orientation of the tissue samples varied, resulting in a different number of pixel scanning the tissue for each sample (Figure 44). Consequently, the data for each tissue sample had a different number of pixels (waveforms) or rows N. The

detailed histopathology results were appended to the Terahertz (THz) data for each pixel. The final dataset contained a number of rows (pixels), containing each row of the waveform 301 data points and the percentage of each tissue type (tumour, fibrous, adipose) present in that particular pixel. A total number of 257 pixels was collected resulting in a matrix of n=257 waveforms by m=301 data points. The dataset contained 115 tumour pixels, 116 fibrous pixels and 26 pure adipose pixels. The tumour pixels predominantly consisted of invasive ductal/no special type carcinoma (N = 92) and invasive lobular carcinoma (N = 19). Most of the tumour pixels contained a low to moderate percentage of tumour cells ranging between 1 – 60% (N = 98). Almost all tumour cells had a background of pure fibrous tissue; only five had a background containing a mix of fibrous and adipose. Most of the fibrous pixels had a high percentage of fibrous cells ranging between 81 – 100% (N = 91). Only 26 of the 257 pixels consisted of pure adipose tissue. A detailed description of the pixel characteristics of the dataset is illustrated in Table 23.

Table 23. Pixel characteristics analysis dataset. A total of 257 pixels were included in the TPI dataset: 115 tumour pixels, 116 fibrous pixels and 26 pure adipose pixels. Figure taken from Grootendorst et al. 2017 (321).

	Tumour					**Fibrous**		**Adipose**
Tissue percentage groups (%)	IDC/ NST	IDC/ NST + DCIS	ILC	No. of pixels	BG	No. of pixels	BG	No. of pixels
81 – 100	3	1		4	F	91	A	26[1]
61 – 80	11	2		13	F	2	A	
41 – 60	22		6	28	F	7	A	
21 – 40	33	1	12	46	F: 43 F/A: 3	3	A	
1 – 20	23		1	24	F: 22 F/A: 2	13	A	
No. of pixels	92	4	19	115		116		26

IDC/NST = invasive ductal/no special type carcinoma; DCIS = ductal carcinoma in situ; ILC = invasive lobular carcinoma; BG = background tissue. In our dataset, the background consisted of fibrous tissue (F), adipose tissue (A), or a mixture of fibrous and adipose tissue (F/A).

1 These pixels contained 100% adipose tissue.

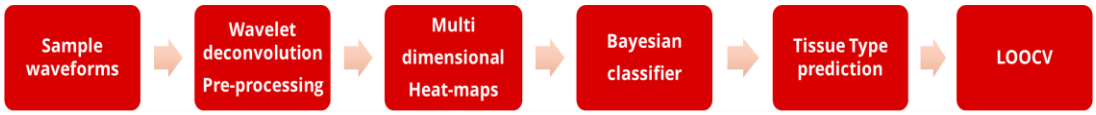
The raw data was therefore highly complex, multidimensional and contained lots of noise. There was a lack of standardised procedures for data acquisition, post processing, image analysis and interpretation of the data generated with this technology.

ii. Statistical Analysis

The analytical pipeline followed in this project is an innovative approach to process TPI THz data, previously analysed using support vector machine, a traditional machine learning technique (341, 343).

The statistical pipeline in this project is therefore divided into three main analyses performed to understand individuals' susceptibility to cancer by exploring the variability on tissue composition with the ultimate aim of identifying discriminating features in imaging data that will allow us to classify breast tissue as malignant or healthy tissue in real-time surgery (Figure 45). The analyses comprising the statistical pipeline were the following: (1) **Gaussian wavelet deconvolution algorithm**, a signal processing algorithm was first developed to reduce the noise and expand the information of the original signals, (2) **Naïve Bayes Classifier**, was applied to find rules or patterns (if any) that were associated with the histopathological composition of the tissue samples and the TPI waveforms using the prior information on the correlated histopathological composition known for each waveform (pixel) and finally (3) **Leave one sample out cross validation (LOOCV),** was applied to predict the histopathological composition of an observation (waveform) while leaving all the rest of waveforms from the same scan of the tissue sample out of the dataset employed for prediction, to ultimately evaluate the prediction accuracy of the Bayesian algorithm on unseen samples.

Figure 45. Methodological approach using an innovative avenue to explore individual susceptibility to cancer using TPI imaging data. The raw sample waveforms are modelled via Gaussian wavelet deconvolution generating multidimensional heat-maps of imaging data. The heat-maps are then used as an input model for the Bayesian classifier that will predict the different tissue types based on the true histopathology values. Finally, the accuracy of the prediction is measured applying leave one sample out cross validation (LOOCV).



Furthermore, to perform the above-described statistical pipeline, the **Gaussian wavelet deconvolution** was applied to the high dimensional dataset which constituted of the TPI waveforms from all 46 breast samples included for analysis (matrix of n=257 waveforms by m=301 data points). Gaussian deconvolution was considered as an adequate signal processing transformation for this data due to the analogy of the TPI impulse function and the derivatives of the Gaussian function. The transformation aimed to reduce the noise and to expand the information of the TPI signal.

The transformation of the original TPI waveform was based on the calculation of Gaussian derivatives of different orders (n = 0, 1, 2, 3, 4) to each of the 301 data points which comprised each signal, thereby producing one heat-map per order of derivatives per pixel (Figure 46, Figure 47). Higher order Gaussian derivatives were not used to avoid potential overfitting.

Figure 46. Gaussian wavelet deconvolution signal transformation applied to the TPI dataset involved Gaussian convolutions of derivatives of the original time series. Below the standard formula of discretised approximations of these derivatives, of order 1 to 4 respectively, is illustrated.

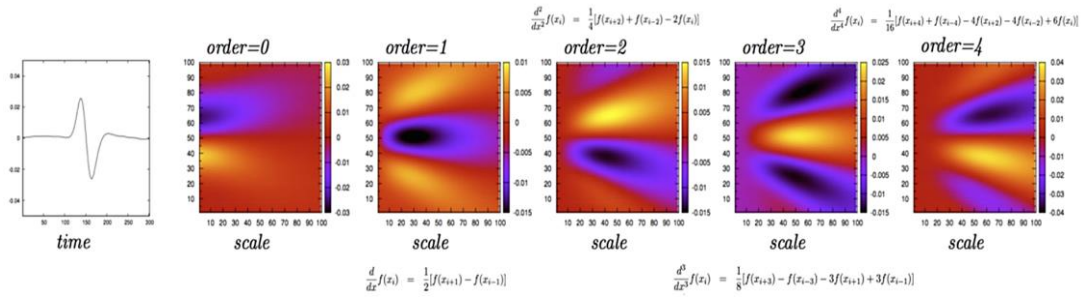
$$\frac{d}{dx} f(xi) = \frac{1}{2} [f(xi+1) - f(xi-1)]$$

$$\frac{d2}{dx2} f(xi) = \frac{1}{4} [f(xi+2) + f(xi-2) - 2f(xi)]$$

$$\frac{d3}{dx3} f(xi) = \frac{1}{8} [f(xi+3) - f(xi-3) - 3f(xi+1) + 3f(xi+1)]$$

$$\frac{d4}{dx4} f(xi) = \frac{1}{16} [f(xi+4) + f(xi-4) - 4f(xi+2) - 4f(xi-2) + 6f(xi)]$$

Figure 47. Example of the heat-maps generated from one TPI waveform. Gaussian derivatives of different orders (n=0, 1,2,3,4).



The five heat-maps per waveform were merged together into a multidimensional heat-map per waveform which also incorporated as corresponding outcome variable the specific tissue content of each of the pixels and was defined as follows: percentage of tumour, percentage of fibrous and percentage of adipose tissue in the pixel (e.g.: %tumour=35, %fibrous=45 and %adipose=20).

A **Naïve Bayes Classifier** algorithm was then applied to the multidimensional heat-maps to discriminate benign from malignant breast tissue (166). The classifier used the known classification of the observations to find the rules or patterns (if any) that link those observations to their classes. These rules were then applied to predict the classes of new observations from a test or validation set. Here, the classifier used the pathologically defined tissue content of each of the observed TPI waveforms of each of the pixels to establish patterns that linked the specific waveforms to the tissue content, so that it could then employ these rules to predict the tissue content in a new breast TPI signal.

The classifier algorithm implemented a probabilistic approach, or model-based, where it was assumed that the observations in each class were generated by a distribution that was specific to that class, given that the distributions were multivariate Gaussian.

This algorithm was preferred among other classification methods given its capacity to rapidly analyse multidimensional dataset with bigger number of covariates or predictors M that number of samples N.

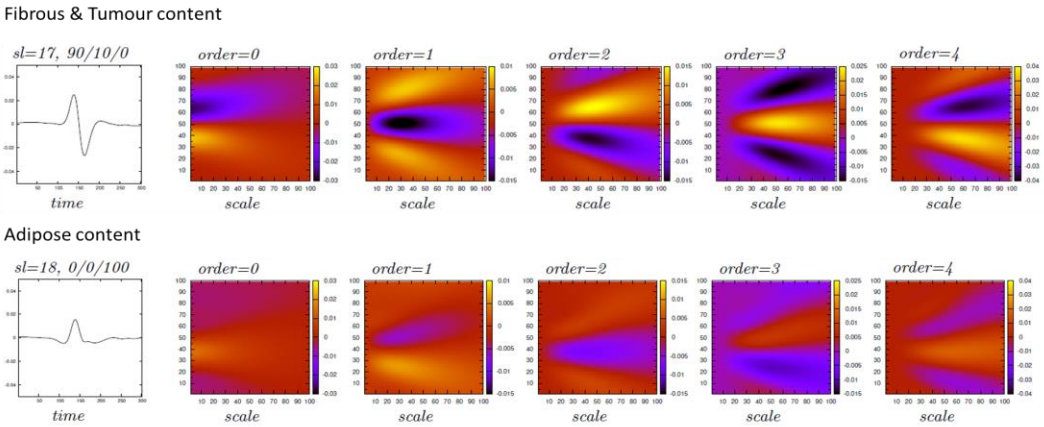
To simplify the nature of the tissue when training the classifier, the classifier was programmed to implement three scenarios that assessed the type of tissue available in the sample in a different manner depending on the particular research question evaluated. **Scenario 1**: pixels were marked as tumour when containing any amount of cancer cells (if %tumour>0 TPI waveform was considered as tumour, irrespective of the content of fibrous or adipose tissue), otherwise the tissue content of the pixel was defined by the highest percentage of fibrous or adipose tissue; **Scenario 2**: pixels were marked as tumour when containing any amount of cancer cells (if %tumour>0 pixel TPI waveform was considered as tumour, independently of the content of fibrous or adipose tissue), otherwise tissue was considered to be "benign". In this scenario, adipose and fibrous tissues were grouped together; **Scenario 3**: pixels were marked based on their tissue content, so the tissue content of each pixel was defined by the highest percentage. Given that the main aim of the project was to discriminate malignant tissue from benign breast tissue, and that benign breast tissue could be subdivided into fibrous and adipose tissue respectively, scenario 1 was the preferred option in the study. The naïve Bayes classifier was employed using both the raw waveforms dataset and the multidimensional heat-maps dataset for each of the three scenarios to assess the best input model for the classifier.

Finally, the performance of the classifier was evaluated using the '**leave one sample out cross validation (LOOCV)**'. The Bayesian classifier was trained individually by leaving out the pixels of a single sample to be classified, and training each classifier with the other samples. The trained classifiers were then applied to the pixels of the sample that was left out. This process was repeated for all the samples. The results were compiled to estimate accuracy, sensitivity, specificity, positive predictive value, and negative predictive value to distinguish malignant from benign tissue. Moreover, leave one pixel out cross validation was also explored to assess the performance of the classifier when considering each pixel independent of the rest of the dataset.

c. Results

A visual examination of the Gaussian derivatives of order 0 to 4, generated with the wavelet deconvolution signal processing algorithm when applied to the waveforms, presented clear differences between the pixels with a high adipose component and the pixels with tumour and fibrous content as illustrated in Figure 48. However, the samples with tumour content did not showed any distinct visual difference with the samples with high fibrous content.

Figure 48. Examples of heat-maps from a waveform generated from tissue with fibrous and tumour content and a heat-map from a waveform from adipose tissue. Clear visual differences exist between both heat-maps for all the Gaussian derivative orders.



The naïve Bayes classifier was applied on the raw waveforms dataset and on the heat-maps multidimensional dataset for the three scenarios, measuring the predictive capabilities of the classifier using both the leave one sample out cross validation and the leave one pixel out cross validation. The classifier showed a better performance when the heat-maps multidimensional dataset was the input model in comparison with the raw waveforms dataset (Figure 49). Scenario 1 which aimed to predict the presence of tumour in any percentage on the tissue sample, even for the lower strata of 1-20 percentage of tumour present, obtained lower sensitivity and higher specificity compared to scenario 2 (Figure 50) and lower predictive scores compared to scenario 3, given the prevalence of the low tumour content pixels on the dataset (Table 23). When applying the leave one pixel out cross validation, the results were better than performing the leave one sample out cross validation.

This was expected, because the leave one pixel out cross validation did not consider any possible correlations between the pixels belonging to the same scanned tissue sample (Figure 50). To replicate the real-time surgery context, the final results were based on the performance of the naïve Bayes classifier when scenario 1 and LOOCV was employed.

Figure 49 presents the final classification results for the 46 breast cancer specimens by tissue type, using both the original waveform and the processed heat-map as the input model of the Bayesian classified when performed using scenario 1 and leave one sample out was applied. The accuracy, sensitivity, specificity, PPV and NPV of the Bayesian classifier for tumour tissue discrimination, performed using the combined Gaussian derivatives of order 0 (normal Gaussian function), 1, 2, 3 and 4 applied to the TPI waveforms, were 69%, 89%, 53%, 60%, 86%, respectively. The classifier considered tumour pixels when the pixel contained any amount of cancer cells. Tumour tissue type scanlines were classified correctly in 89% of the cases using the heat-map derivatives from the time domain pulse in the Bayesian classifier with an accuracy of 0.7 and sensitivity of 0.9. The adipose tissue scanlines were accurately discriminated from tumour/fibrous tissue using the heat-map derivatives with an accuracy of 0.95. Fibrous tissue scanlines were incorrectly misclassified as tumour in 65% of the cases.

Figure 49. Classification results for the 46 breast cancer specimens by tissue type, using both the original waveform and the processed heat-map as our input model of the Bayesian classified when performed using scenario 1 and leave one sample out was applied. Tumour and Adipose scanlines are correctly classified in more than 89 % of the cases.

Input File		**Original Waveform**			**Heatmap**	
LOOCV Coding		1			1	
True Class	Tumour	Fibrous	Adipose	Tumour	Fibrous	Adipose
Percentage of the class on the sample	0.437	0.403	0.160	0.437	0.403	0.160
Predict_Tumour	85%	68%	0%	89%	65%	0%
Predict_Fibrous	15%	24%	7%	11%	26%	7%
Predict_Adipose	0%	8%	93%	0%	8%	93%
Accuracy	0.67	0.62	0.95	0.69	0.64	0.95
Sensitivity	0.85	0.24	0.93	0.89	0.26	0.93
Specificity	0.51	0.88	0.96	0.53	0.90	0.96
PPV	0.58	0.56	0.81	0.60	0.64	0.81
NPV	0.82	0.63	0.99	0.86	0.64	0.99

Figure 50. Tumour tissue classification results when using Bayes classifier on heat-maps dataset comparing scenario 1 and 2 performance and leave-one-sample-out and leave-one-pixel-out cross validation.

LOOCV	**1**	**1**	**2**	**2**
Coding	1	2	1	2
Accuracy*	0.69	0.67	0.70	0.67
Sensitivity*	0.89	0.94	0.91	0.96
Specificity*	0.53	0.45	0.53	0.45
PPV*	0.60	0.57	0.60	0.58
NPV*	0.86	0.90	0.89	0.93

LOOCV 1= Leave-one-sample-out cross validation 2= Leave-one-pixel-out-cross validation
Coding 1= Scenario 1 2= Scenario 2
***** = All the parameters are detection characteristics calculated for tumour using the heat-maps dataset

Figure 51 presents the tissue type content of the waveforms in the dataset. The classification by tissue type by waveform (pixel) showed that of the 115 tumour pixels, 89% were classified correctly as tumour and 15 were misclassified as fibrous. All misclassified pixels contained ≤60% tumour cells and included a high percentage of the ILC samples in the dataset (6/19) (Figure 52). Sixty-six of the 142 benign pixels were wrongly classified as tumour; 64 of these were 81 – 100% fibrous pixels, only two 1 – 80% fibrous pixels were misclassified. The scanlines containing 81-100% of fibrous are misclassified as tumour in a 70% of the cases. All pure adipose pixels were correctly classified (Figure 53).

Figure 51. Types of pixels (waveforms) in the dataset (n=257).

No. of Benign pixels	142
No. of Adipose	26
No. of 81-100% Fibrous	91
No. of 1-80% Fibrous	25
No. of Tumour pixels	115

Figure 52. Classification of the tumour content for the pixels on the dataset based on the final results of the classifier (Figure 49).

No. of Tumour correctly classified	100
No. of Tumour misclassified	15
IDC /NST misclassified	8
NST with DCIS misclassified	1
ILC misclassified	6

Figure 53. Classification of the benign content for the pixels on the dataset based on the final results of the classifier (Figure 49).

Total of Benign correctly classified	76
No. of pure Adipose misclassified as tumour	0
No. of Benign misclassified as tumour	66
No. of 81-100% Fibrous misclassified as tumour	64
No. of 1- 80 % Fibrous misclassified as tumour	2

d. Discussion

To investigate individuals' susceptibility to cancer by examining the heterogeneity inherent to cancer, I aimed to distinguish variability at cellular level by differentiating tissue composition (tumour, fibrous and adipose material) in a translational setting, using imaging data generated from ex vivo breast tissue samples. The project explored the capability of a naïve Bayes classification algorithm to discriminate malignant from benign breast tissue for a real-time surgery utilisation with the ultimate aim of improving breast cancer surgical outcomes. The innovative statistical approach followed in the project, indicated that wavelet deconvolved pre-processed imaging data when analysed with the Bayesian classifier can accurately distinguish malignant from benign tissue in an *ex vivo* setting, obtaining high sensitivity and low specificity. More specifically, I observed that tumour and adipose tissue were correctly classified in most of the cases (89%), however high density fibrous was incorrectly classified as tumour and adipose in 73% of the cases.

Moreover, the results suggested that supervised clustering methods can effectively classify multidimensional complex biological data extracting the hidden patterns to accurately predict new observations in the data. Further optimisation of these new approaches could help improving diagnosis in translational and clinical settings.

Malignant and benign tissue

The raw waveforms and the processed heat-maps visually presented distinct differences between the images produced by the adipose tissue in comparison with

the plots produced by tumour and fibrous tissue samples (Figure 44 and Figure 48). This confirmed the results by *Ashworth et al*. observed in a small study where the TPI generated THz impulse functions from fibrous tissue and breast cancer showed similar features while the impulse functions from adipose tissue were dissimilar (344). The classifier results confirmed these observations being adipose the tissue type with higher classification performance presenting values over 90% for accuracy, sensitivity and specificity. Tumour tissue was correctly predicted in 89% of the cases, which was reflected in a high sensitivity score (89%) and percentage of true positive cases identified by the classifier. This suggest the potential clinical applicability of this method to identifying tumour cells close to a resection margin intraoperatively, with the subsequent possibility to improve the adequate excision of the margins reducing re-operation rates, however the potential applicability of the classifier is conditioned on the improvement on the specificity score. Tumour specificity showed moderate results (53%) as a consequence of the high number of fibrous samples misclassified as tumour (65% of the cases). Exploring the 257 pixels generated by scanning the 46 BC samples, the tumour samples contained predominantly low-to-moderate tumour cell percentages, which replicates the tissue composition of the resection border of breast samples from patients with positive margins. Moreover, the benign dataset contained a high representation of 81-100% high dense fibrous (91/142 pixels), from which 64 pixels were misclassified as tumour. This suggests that tumour cells might present similar water content and cell density as high dense fibrous material when measured with TPI terahertz technology.

Imaging data analysis

The proposed statistical pipeline to process and analyse the TPI imaging data showed good classification capabilities to discriminate tumour samples, however it only had moderate capability of discriminating fibrous. The naïve Bayes classifier explored the information contained in the raw waveforms given that the Gaussian derivative heat-maps were performed based on the similarity of the distribution of the raw data to the Gaussian distribution. This mathematical approach given its

capability to rapidly analyse multidimensional matrixes, then avoided the evaluation of standard heuristic parameters predefined for imaging data such as amplitude or peak to peak, commonly used to analyse imaging data, as for example utilised in previous analysis of TPI imaging data, obtaining similar results with a faster performance (SVM: sensitivity (86%), specificity (66%)) (321). Moreover, the sensitivity and specificity obtained by the classifier (89%, 53%) were comparable to the performance of IMAs techniques currently used for intraoperatively assess tumour margins during BCS, given that there is broad variability between the current IMASs as reported by a meta-analysis such as specimen radiography (53%, 84%), ultrasound imaging (59%, 81%), radiofrequency spectroscopy (71%, 68%) or frozen section analysis (68%, 96%) (337). Most of the techniques explored in the meta-analysis presented lower sensitivity than the Bayesian classifier, which indicates that these IMAs are not able to classify tumour tissue as tumour with the same accuracy as the Bayesian classifier. However, these IMAs presented better specificity, which means they are able to detect false negatives more accurately than the Bayesian classifier. They will identify fibrous tissue more accurately. Moreover, given that frozen section analysis is considered the gold standard, the Bayesian classifier needs further development to improve the specificity score.

Future steps

Based on the results of the present study, the TPI imaging data, when analysed with supervised clustering methods, has the ability to assess intraoperatively tumour margins quite effectively given the sensitivity obtained in the project. However, the performance in terms of specificity needs to improve to meet the needs of this specific diagnostic setting. Therefore, different steps can be followed to improve the specificity of the classifier: (i) explore the predictive capability of each of the Gaussian derivatives independently, (ii) optimise the Bayesian classifier to improve its performance, and (iii) acquire a new dataset increasing the number of BC tissue samples including data from diverse pathologically well-defined breast tumour types and multiple measurements per BC patient to be able to assess inter and

intraindividual variability. The latter was not possible in this project given the relative small dataset.

Strengths and limitations

The major strength of these analyses lies in the innovative avenue to study individuals' susceptibility by examining the heterogeneity inherent to cancer tissue in a translational imaging setting. The dataset analysed was generated in the first-in-human, single centre pilot study to evaluate the ability of a handheld TPI probe to discriminate benign from malignant breast tissue in an ex vivo setting conducted in KHP. Forty-six freshly excised breast samples with low to moderate tumour content from various BC types generated 257 THz pixels which were included in the analysis. However, a dataset more representative of the different tissue composition, including more low density fibrous, more adipose samples and diverse BC molecular subtypes, together with multiple measurements per patient would better reflect the heterogeneity inherent to the disease and would replicate the tissue composition variability on the resection border of breast samples in a live surgery setting.

Conclusion

The results of the Bayesian classifier adapted in this study illustrate how an individuals' susceptibility to disease can be explored examining the heterogeneity inherent to cancer by investigating tissue composition (tumour, fibrous and adipose material) in a translational setting using imaging data. The naïve Bayes classification can effectively discriminate malignant from benign breast tissue using TPI imaging data, however presents a low specificity. This suggests the potential clinical applicability of this method to improve the adequate excision of the margins in BCS surgery if the classifier is optimised to improve the specificity score.

Thus, based on the results of this exploratory exercise, it can be concluded that supervised clustering approaches present the capability to classify multidimensional complex biological data whilst extracting hidden patterns to accurately predict new

observations in the data. In the future, these innovative approaches could help improving diagnosis and prognosis in translational and clinical settings.

Chapter VI: Conclusion and future directions

Cancer burden continues to increase in an aging population. It has evolved into a chronic condition in high income countries, whilst in developing countries early detection is still a critical milestone in the cancer pathway (14). Moreover, cancer is characterised by a multifactorial aetiology with an intricate gene-environment interplay (228), which implies the involvement of diverse biological pathways, including distinct genetic, molecular and clinical events between patients (4, 29). Biological variability is thus inherent to the disease and population susceptibility depends on a tight reciprocity between biological components modulated by genetics, life-style behaviour aspects and environmental factors (39). Due to this complexity, single biomarkers and single clinical symptoms seem insufficient to understand the disease and tailor patient treatments. Hence, a new research avenue towards the pursuit of personalized medicine has arisen. Novel systematic efforts exploring the vast amount of cancer data, acquired with new technologies, in combination with the implementation of the exposome concept and advanced statistical methods, have the potential to build efficient stratification models for predictive health that could be implemented in clinical practice (345).

Hence, this thesis was written with the ultimately aim of acquiring a comprehensive perspective of the population susceptibility to cancer and variability of the disease whilst investigating new statistical approaches using multidimensional cancer datasets, in the pursuit of an accurate and effective stratified medicine for cancer. Therefore, three different projects were performed, two of them exploring population susceptibility to cancer in a clinical setting and one exploring individuals' susceptibility to cancer in a translational imaging setting. These projects helped to investigate and generate hypotheses and new research ideas for future work.

Thus, the next section provides a summary of the conclusions for each project, followed by a description of new research avenues to continue these projects in the near future.

The chapter *Population susceptibility to cancer* included Project A - the blood exposome in AMORIS and Project B - metabolic profiles in AMORIS. The projects aimed to explore multiple standard of care serum markers available in AMORIS in relation to cancer diagnosis and mortality, firstly by understanding the role of this subset of serum markers in the blood exposome and secondly, by characterizing the potential of these established markers as markers of cancer susceptibility.

The practical framework established in this thesis to study a subset of markers of the blood exposome in AMORIS showed that internal biological markers were strongly correlated with markers of related body functions and biological pathways. Simultaneously novel networks within markers of *a priori* unrelated functions, were also observed. Thus, this complex synergy between markers, explained the reason why the variance of the data was distributed along the different biomarkers, without a group of markers that could account for all the variability of the data. This exploratory analysis demonstrated the need of systemic approaches involving multiple markers capable of evaluating the internal biological environment. Moreover, the interactions observed with the panel external (2), specific external markers (2) and the internal molecules (21) included in this study illustrate the interaction between the environmental, life style factors and the internal biological component in human populations for this small subset of the exposome. The results of this exploratory analysis of a subset of the exposome highlight the importance of the implementation of practical frameworks to characterise the exposome that will enable us to unlock the heterogeneity in a given population, which emphasises the relevance of the assessment of the exposome when measuring exposure impact and health outcomes.

The statistical pipeline developed in this thesis to characterise the potential of blood markers as markers of cancer susceptibility in AMORIS demonstrated that routinely collected serum markers, when assembled into meaningful metabolic components, have the potential to stratify the population for cancer risk, cancer mortality and overall mortality. The metabolic profiles represented by lipid molecules, the

enzymes ALT, AST and GGT and the iron metabolites classified individuals at higher risk of cancer and mortality, suggesting a key role of these markers in the internal body homeostasis. Moreover, these results suggested the potential use of standard serum markers as markers of cancer susceptibility in early detection of cancer. The optimised statistical analysis revealed a promising role of data reduction methods in risk stratification modelling for biomedical data, which needs to be established in larger datasets including broad panels of markers and lifestyle and environmental factors such as the EPIC cohort - given the moderate prediction scores shown by the LCA metabolic profiles in comparison to the single established markers (Table 22)Therefore, the project highlights that data driven efforts using multiple markers and statistical advanced technologies are necessary to understand disease and population susceptibility to disease, and could help to implement better stratification programs for prevention and early detection in cancer.

Given the observed stratification capabilities of the metabolic profiles for cancer risk and mortality in AMORIS, a nested case-control study could be conducted in the same cohort to establish whether these profiles could be utilised to predict cancer risk and mortality in the Swedish population. Moreover, to assess the external validity of the metabolic profiles as markers of cancer susceptibility, a validation study is currently being conducted in a different population characterised by diverse life-style, demographics and genetics, using the American National Health and Nutrition Examination Survey (NHANES) cohort data (outside the scope of this PhD thesis). This validation study includes the final panel of biomarkers performed in AMORIS as the crude model, together with other models including different standard serum markers, such as red blood count, diverse types of white blood cells, haemoglobin or immunoglobulins available in NHANES. The results of the LCA analysis for the crude model have already identified four main metabolic classes in the population defined by similar markers as the serum markers identified in the AMORIS project, which suggests the external validity of the metabolic profiles to extract the population heterogeneity as well as its potential use as markers of cancer susceptibility.

Overall, future work should aim to investigate the exposome in relation to disease and mortality to improve prediction of complex, multifactorial diseases, such as cancer, which in turn could lead to better public health strategies to prevent chronic diseases. However, longitudinal studies to account for exposome temporal variation with measurements at different relevant life time points will be necessary to understand the complex interaction between environment, biology and disease in human populations.

Following the findings of this thesis, I have already been involved in the development of two new projects in AMORIS to better understand the exposome and its association with disease. First, the twin study, a small subcohort of twins with monozygotic and dizygotic status available in AMORIS, will be utilised. The aim of this project will be to quantify the proportion of the environmental and the genetic contribution of the serum markers to cancer. The second project will consist on the utilisation of job matrixes available for the Swedish population during the same period as the CALAB blood analyses were taken. The link between the job matrixes available and the serum markers from the same period will allow to identify specific environmental exposures for the participants of the AMORIS dataset. The combined information will then favour a better assessment of the exposome in relation to cancer. Moreover, there will be a possibility to make use of newly developed data reduction techniques of Professor Ton Coolen to further explore population heterogeneity (166, 346).

The chapter *Individual susceptibility to cancer* included Project C - Discrimination of breast tumour tissue. The project aimed to evaluate how imaging data produced by a terahertz probe can distinguish between different breast tissue samples, with the ultimate aim of using this probe intraoperatively in breast-conserving surgery to predict positive tumour margins. The study illustrated how individuals' susceptibility to disease can be explored examining the heterogeneity inherent to cancer present on tissue composition (tumour, fibrous and adipose material) using ex vivo breast

cancer samples in an imaging setting. The two-step statistical approach (Gaussian wavelet deconvolution of the raw imaging data followed by the naïve Bayes classifier) showed that it could accurately discriminate malignant from benign tissue in an *ex vivo* setting, obtaining high sensitivity and moderate specificity, classifying correctly tumour and adipose tissue in 89% of the cases. The ultimately aim of the classifier was to be implemented in BCS surgery to improve adequate excision of the tumour margins avoiding re-operations in patients. This clinical application can be achieved in future with an improvement on the specificity score of the classifier. For that purpose, a follow up project is currently being conducted exploring the predictive capabilities of the different Gaussian derivatives and utilising an optimized Bayesian classifier to improve the sensitivity and specificity of the classifier to discriminate malignant breast tissue (346). Thus, supervised clustering approaches present the capability to classify multidimensional complex biological and accurately predict new observations in the data which could help improving diagnosis and prognosis in translational and clinical settings.

Overall, the projects in this thesis highlight the importance of data driven approaches in the assessment of multifactorial diseases, such as cancer, when supported by robust statistical analysis. Moreover, the results suggest that the inherent heterogeneity present in the population plays an important role in the susceptibility to cancer, which needs to be taken into consideration to develop efficient stratification strategies for prevention and early cancer detection.

References

1. Welch DR. Tumor Heterogeneity—A 'Contemporary Concept' Founded on Historical Insights and Predictions. Cancer Res. 2016;76(1):4-6.
2. Prensner JR, Rubin MA, Wei JT, Chinnaiyan AM. Beyond PSA: The next generation of prostate cancer biomarkers. Science translational medicine. 2012;4(127):127rv3-rv3. PubMed PMID: PMC3799996.
3. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000 Jan 7;100(1):57-70. PubMed PMID: 10647931.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 04;144(5):646-74. PubMed PMID: 21376230. Epub 2011/03/08. eng.
5. Mabert K, Cojoc M, Peitzsch C, Kurth I, Souchelnytskyi S, Dubrovska A. Cancer biomarker discovery: Current status and future perspectives. Int J Radiat Biol. 2014 May 12:1-19. PubMed PMID: 24524284. Epub 2014/02/15. Eng.
6. Nair M, Sandhu SS, Sharma AK. Prognostic and Predictive Biomarkers in Cancer. Curr Cancer Drug Targets. 2014 May 6. PubMed PMID: 24807144. Epub 2014/05/09. Eng.
7. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nature clinical practice Oncology. 2008 Oct;5(10):588-99. PubMed PMID: 18695711. Epub 2008/08/13. eng.
8. Brooks JD. Translational genomics: the challenge of developing cancer biomarkers. Genome Res. 2012 Feb;22(2):183-7. PubMed PMID: 22301132. Pubmed Central PMCID: 3266026.
9. Diamandis EP. Biomarker validation is still the bottleneck in biomarker research. J Intern Med. 2012 Dec;272(6):620. PubMed PMID: 22905877.
10. Zhang Y. News & Views: Bring on the Biomarkers—It's Time for the "Big Science" Approach. Clin Chem. 2011 June 1, 2011;57(6):928-9.
11. Andreopoulos B, An A, Wang X, Schroeder M. A roadmap of clustering algorithms: finding a match for a biomedical application. Briefings in Bioinformatics. 2009 May 1, 2009;10(3):297-314.
12. Zhao Y, Karypis G. Data clustering in life sciences. Mol Biotechnol. 2005 Sep;31(1):55-80. PubMed PMID: 16118415. Epub 2005/08/25. eng.
13. Garrett ES, Zeger SL. Latent class model diagnosis. Biometrics. 2000 Dec;56(4):1055-67. PubMed PMID: 11129461. Epub 2000/12/29. eng.
14. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011 Mar-Apr;61(2):69-90. PubMed PMID: 21296855. Epub 2011/02/08. eng.
15. Global Burden of Disease Cancer C, Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol. 2017 Apr 1;3(4):524-48. PubMed PMID: 27918777.

16. Global Burden of Disease Cancer C, Fitzmaurice C, Dicker D, Pain A, Hamavid H, Moradi-Lakeh M, et al. The Global Burden of Cancer 2013. JAMA Oncol. 2015 Jul;1(4):505-27. PubMed PMID: 26181261. Pubmed Central PMCID: 4500822.
17. Restifo NP, Dudley ME, Rosenberg SA. Adoptive immunotherapy for cancer: harnessing the T cell response. Nature Reviews Immunology. 2012 03/22/online;12(4):269. PubMed PMID: 22437939.
18. Bertram JS. The molecular biology of cancer. Mol Aspects Med. 2000 Dec;21(6):167-223. PubMed PMID: 11173079. Epub 2001/02/15. eng.
19. Hunter T. Oncoprotein Networks. Cell. 1997 2/7/;88(3):333-46.
20. Easty D, Gallagher W, Bennett DC. Protein tyrosine phosphatases, new targets for cancer therapy. Curr Cancer Drug Targets. 2006 Sep;6(6):519-32. PubMed PMID: 17017875. Epub 2006/10/05. eng.
21. Motiwala T, Jacob ST. Role of Protein Tyrosine Phosphatases in Cancer. Prog Nucleic Acid Res Mol Biol. 2006;81:297-329. PubMed PMID: PMC3077959. Pubmed Central PMCID: 3077959.
22. Polakis P. Wnt Signaling in Cancer. Cold Spring Harbor Perspectives in Biology. 2012 May 01;4(5):a008052. PubMed PMID: PMC3331705. Pubmed Central PMCID: 3331705.
23. Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. Oncogene. 2016 09/12/online;36(11):1461. PubMed PMID: 27617575. Pubmed Central PMCID: 5357762.
24. Stemmler MP. Cadherins in development and cancer. Molecular bioSystems. 2008 Aug;4(8):835-50. PubMed PMID: 18633485. Epub 2008/07/18. eng.
25. Singhai R, Patil VW, Jaiswal SR, Patil SD, Tayade MB, Patil AV. E-Cadherin as a diagnostic biomarker in breast cancer. N Am J Med Sci. 2011 May;3(5):227-33. PubMed PMID: 22558599. Pubmed Central PMCID: 3337742.
26. Ridley AJ. Rho proteins and cancer. Breast Cancer Res Treat. 2004 Mar;84(1):13-9. PubMed PMID: 14999150. Epub 2004/03/05. eng.
27. Fernández-Medarde A, Santos E. Ras in Cancer and Developmental Diseases. Genes & Cancer. 2011;2(3):344-58. PubMed PMID: PMC3128640.
28. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signalling pathways in cancer. Oncogene. 2007 May 14;26(22):3279-90. PubMed PMID: 17496922. Epub 2007/05/15. eng.
29. Pavlova Natalya N, Thompson Craig B. The Emerging Hallmarks of Cancer Metabolism. Cell Metabolism. 2016 2016/01/12/;23(1):27-47. PubMed PMID: 26771115. Pubmed Central PMCID: 4715268.
30. Heppner GH. Tumor Heterogeneity. Cancer Res. 1984;44(6):2259-65.
31. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013 09/18/online;501(7467):338. PubMed PMID: 24048066.
32. Dai X, Xiang L, Li T, Bai Z. Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes. Journal of Cancer. 2016;7(10):1281-94. PubMed PMID: 27390604. Pubmed Central PMCID: PMC4934037. Epub 2016/07/09. eng.
33. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer. 2013 01/08/online;108(3):479. PubMed PMID: 23299535. Pubmed Central PMCID: 3593543.

34. Sheffield BS. Immunohistochemistry as a Practical Tool in Molecular Pathology. Arch Pathol Lab Med. 2016 Aug;140(8):766-9. PubMed PMID: 27472235. Epub 2016/07/30. eng.
35. Viale G. The current state of breast cancer classification. Ann Oncol. 2012;23(suppl_10):x207-x10.
36. Perou CM. Molecular stratification of triple-negative breast cancers. Oncologist. 2010;15 Suppl 5:39-48. PubMed PMID: 21138954. Epub 2010/12/16. eng.
37. The Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012 09/23/online;490:61.
38. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346-52.
39. Kolonel LN, Altshuler D, Henderson BE. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. Nature Reviews Cancer. 2004 07/01/online;4:519.
40. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The Diploid Genome Sequence of an Individual Human. PLoS Biol. 2007;5(10):e254.
41. Bernstein H, Bernstein C. Evolutionary Origin of Recombination during Meiosis. Bioscience. 2010;60(7):498-505.
42. Ewens WJ, Lessard S. On the interpretation and relevance of the Fundamental Theorem of Natural Selection. Theor Popul Biol. 2015 Sep;104:59-67. PubMed PMID: 26220589. Epub 2015/07/30. eng.
43. The International HapMap C. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 09/02/online;467:52.
44. Ring HZ, Kwok PY, Cotton RG. Human Variome Project: an international collaboration to catalogue human genetic variation. Pharmacogenomics. 2006 Oct;7(7):969-72. PubMed PMID: 17054407.
45. International Human Genome Sequencing C. Initial sequencing and analysis of the human genome. Nature. 2001 02/15/online;409:860.
46. The International SNPMWG. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001 02/15/online;409:928.
47. Kinzler KW, Vogelstein B. Cancer-susceptibility genes. Gatekeepers and caretakers. Nature. 1997 Apr 24;386(6627):761, 3. PubMed PMID: 9126728. Epub 1997/04/24. eng.
48. Churpek JE, Walsh T, Zheng Y, Moton Z, Thornton AM, Lee MK, et al. Inherited predisposition to breast cancer among African American women. Breast Cancer Res Treat. 2015 Jan;149(1):31-9. PubMed PMID: 25428789. Pubmed Central PMCID: PMC4298662. Epub 2014/11/28. eng.
49. Hodgson S. Mechanisms of inherited cancer susceptibility. Journal of Zhejiang University Science B. 2008 11/13/received
12/04/accepted;9(1):1-4. PubMed PMID: PMC2170461.
50. Beggs AD, Hodgson SV. Genomics and breast cancer: the different levels of inherited susceptibility. Europ J Hum Genet. 2009 12/17;17(7):855-6. PubMed PMID: PMC2986490.

51. Yang Q, Khoury MJ, Friedman J, Little J, Flanders WD. How many genes underlie the occurrence of common complex diseases in the population? Int J Epidemiol. 2005 Oct;34(5):1129-37. PubMed PMID: 16043441. Epub 2005/07/27. eng.
52. Jing L, Su L, Ring BZ. Ethnic background and genetic variation in the evaluation of cancer risk: a systematic review. PLoS ONE. 2014;9(6):e97522. PubMed PMID: 24901479. Pubmed Central PMCID: PMC4046957. Epub 2014/06/06. eng.
53. Vieira PC, Burbano RM, Fernandes DC, Montenegro RC, Dos Santos SE, Sortica VA, et al. Population stratification effect on cancer susceptibility in an admixed population from Brazilian Amazon. Anticancer Res. 2015 Apr;35(4):2009-14. PubMed PMID: 25862854. Epub 2015/04/12. eng.
54. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A. 2006 08/31 05/24/received;103(38):14068-73. PubMed PMID: PMC1599913.
55. Cook LS, Goldoft M, Schwartz SM, Weiss NS. Incidence of adenocarcinoma of the prostate in Asian immigrants to the United States and their descendants. J Urol. 1999 Jan;161(1):152-5. PubMed PMID: WOS:000077377800044. English.
56. Camacho-Rivera M, Kalwar T, Sanmugarajah J, Shapira I, Taioli E. Heterogeneity of breast cancer clinical characteristics and outcome in US black women--effect of place of birth. Breast J. 2014 Sep-Oct;20(5):489-95. PubMed PMID: 25041223. Epub 2014/07/22. eng.
57. Baquet CR, Horm JW, Gibbs T, Greenwald P. Socioeconomic Factors and Cancer Incidence Among Blacks and Whites. JNCI: Journal of the National Cancer Institute. 1991 Apr 17;83(8):551-7. PubMed PMID: WOS:A1991FG08600011. English.
58. Cook MB, Dawsey SM, Freedman ND, Inskip PD, Wichner SM, Quraishi SM, et al. Sex disparities in cancer incidence by period and age. Cancer Epidemiol Biomarkers Prev. 2009 Apr;18(4):1174-82. PubMed PMID: 19293308. Pubmed Central PMCID: PMC2793271. Epub 2009/03/19. eng.
59. Cook MB, McGlynn KA, Devesa SS, Freedman ND, Anderson WF. Sex disparities in cancer mortality and survival. Cancer Epidemiol Biomarkers Prev. 2011 Aug;20(8):1629-37. PubMed PMID: 21750167. Pubmed Central PMCID: PMC3153584. Epub 2011/07/14. eng.
60. White MC, Holman DM, Boehm JE, Peipins LA, Grossman M, Henley SJ. Age and Cancer Risk: A Potentially Modifiable Relationship. Am J Prev Med. 2014;46(3 0 1):S7-15. PubMed PMID: PMC4544764.
61. Courtenay WH. Constructions of masculinity and their influence on men's well-being: a theory of gender and health. Soc Sci Med. 2000 2000/05/16/;50(10):1385-401.
62. Bergman MM, Scott J. Young adolescents' wellbeing and health-risk behaviours: Gender and socio-economic differences. J Adolesc. 2001;24(2):183-97.
63. Willett WC. Balancing life-style and genomics research for disease prevention. Science. 2002 Apr 26;296(5568):695-8. PubMed PMID: 11976443. Epub 2002/04/27. eng.

64. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000 Jul 13;343(2):78-85. PubMed PMID: 10891514. Epub 2000/07/13. eng.
65. Claude J, Kunze E, Frentzel-Beyme R, Paczkowski K, Schneider J, Schubert H. Life-style and occupational risk factors in cancer of the lower urinary tract. Am J Epidemiol. 1986 Oct;124(4):578-89. PubMed PMID: 3752052.
66. Khan N, Afaq F, Mukhtar H. Lifestyle as risk factor for cancer: Evidence from human studies. Cancer Lett. 2010 Jul 28;293(2):133-43. PubMed PMID: 20080335. Pubmed Central PMCID: 2991099.
67. Wakai K, Inoue M, Mizoue T, Tanaka K, Tsuji I, Nagata C, et al. Tobacco smoking and lung cancer risk: an evaluation based on a systematic review of epidemiological evidence among the Japanese population. Jpn J Clin Oncol. 2006 May;36(5):309-24. PubMed PMID: 16735374. Epub 2006/06/01. eng.
68. Platz EA, Willett WC, Colditz GA, Rimm EB, Spiegelman D, Giovannucci E. Proportion of colon cancer risk that might be preventable in a cohort of middle-aged US men. Cancer Causes Control. 2000 Aug;11(7):579-88. PubMed PMID: 10977102. Epub 2000/09/08. eng.
69. Lauby-Secretan B, Scoccianti C, Loomis D, Grosse Y, Bianchini F, Straif K. Body fatness and cancer—viewpoint of the IARC Working Group. New Engl J Med. 2016 Aug 25;375(8):794-8. PubMed PMID: WOS:000381799300019. English.
70. Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. Lancet. 2008 Feb 16;371(9612):569-78. PubMed PMID: 18280327.
71. Renehan AG, Soerjomataram I, Leitzmann MF. Interpreting the epidemiological evidence linking obesity and cancer: A framework for population-attributable risk estimations in Europe. Eur J Cancer. 2010 Sep;46(14):2581-92. PubMed PMID: 20843487.
72. Renehan AG, Soerjomataram I, Tyson M, Egger M, Zwahlen M, Coebergh JW, et al. Incident cancer burden attributable to excess body mass index in 30 European countries. Int J Cancer. 2010 Feb 1;126(3):692-702. PubMed PMID: 19645011.
73. Renehan AG, Flood A, Adams KF, Olden M, Hollenbeck AR, Cross AJ, et al. Body Mass Index at Different Adult Ages, Weight Change, and Colorectal Cancer Risk in the National Institutes of Health-AARP Cohort. Am J Epidemiol. 2012 Dec 15;176(12):1130-40. PubMed PMID: WOS:000312634900009. English.
74. Bastide NM, Pierre FH, Corpet DE. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. Cancer Prevention Research. 2011 Feb;4(2):177-84. PubMed PMID: 21209396. English.
75. Jakszyn P, Lujan-Barroso L, Agudo A, Bueno-de-Mesquita HB, Molina E, Sanchez MJ, et al. Meat and heme iron intake and esophageal adenocarcinoma in the European Prospective Investigation into Cancer and Nutrition study. Int J Cancer. 2013 Dec 1;133(11):2744-50. PubMed PMID: 23728954. Epub 2013/06/04. eng.
76. Gilsing AM, Fransen F, de Kok TM, Goldbohm AR, Schouten LJ, de Bruine AP, et al. Dietary heme iron and the risk of colorectal cancer with specific mutations in

KRAS and APC. Carcinogenesis. 2013 Dec;34(12):2757-66. PubMed PMID: 23983135. Epub 2013/08/29. eng.
77. Aykan NF. Red Meat and Colorectal Cancer. Oncology Reviews. 2015 12/28
08/15/received
11/26/revised
12/17/accepted;9(1):288. PubMed PMID: PMC4698595.
78. Cross AJ, Leitzmann MF, Gail MH, Hollenbeck AR, Schatzkin A, Sinha R. A Prospective Study of Red and Processed Meat Intake in Relation to Cancer Risk. PLOS Medicine. 2007;4(12):e325.
79. Terry P, Giovannucci E, Michels KB, Bergkvist L, Hansen H, Holmberg L, et al. Fruit, Vegetables, Dietary Fiber, and Risk of Colorectal Cancer. JNCI: Journal of the National Cancer Institute. 2001;93(7):525-33.
80. Makarem N, Nicholson JM, Bandera EV, McKeown NM, Parekh N. Consumption of whole grains and cereal fiber in relation to cancer risk: a systematic review of longitudinal studies. Nutr Rev. 2016 05/04
01/05/received
10/06/revised
11/03/accepted;74(6):353-73. PubMed PMID: PMC4892300.
81. LoConte NK, Brewster AM, Kaur JS, Merrill JK, Alberg AJ. Alcohol and Cancer: A Statement of the American Society of Clinical Oncology. J Clin Oncol. 2017 Nov 07;0(0):JCO.2017.76.1155. PubMed PMID: 29112463.
82. Clegg LX, Reichman ME, Miller BA, Hankey BF, Singh GK, Lin YD, et al. Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results: National Longitudinal Mortality Study. Cancer causes & control : CCC. 2009 11/12;20(4):10.1007/s10552-008-9256-0. PubMed PMID: PMC2711979.
83. Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N, et al. Trends and socioeconomic inequalities in cancer survival in England and Wales up to 2001. Br J Cancer. 2004 03/09/online;90:1367.
84. Danaei G, Vander Hoorn S, Lopez AD, Murray CJL, Ezzati M. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. The Lancet. 2005 2005/11/19/;366(9499):1784-93.
85. Gathani T, Ali R, Balkwill A, Green J, Reeves G, Beral V, et al. Ethnic differences in breast cancer incidence in England are due to differences in known risk factors for the disease: prospective study. Br J Cancer. 2013 10/29/online;110:224.
86. Forouzanfar MH, Alexander L, Anderson HR, Bachman VF, Biryukov S, Brauer M, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015 Dec 05;386(10010):2287-323. PubMed PMID: 26364544. Pubmed Central PMCID: PMC4685753. Epub 2015/09/15. eng.
87. Rappaport SM. Discovering environmental causes of disease. J Epidemiol Community Health. 2012 Feb;66(2):99-102. PubMed PMID: 22199396. Epub 2011/12/27. eng.

88. Sears ME, Genuis SJ. Environmental Determinants of Chronic Disease and Medical Approaches: Recognition, Avoidance, Supportive Therapy, and Detoxification. Journal of Environmental and Public Health. 2012 01/19 07/16/received 10/19/accepted;2012:356798. PubMed PMID: PMC3270432. Pubmed Central PMCID: 3270432.
89. Trosko JE, Upham BL. The emperor wears no clothes in the field of carcinogen risk assessment: ignored concepts in cancer risk assessment. Mutagenesis. 2005;20(2):81-92.
90. Goodson WH, 3rd, Lowe L, Carpenter DO, Gilbertson M, Manaf Ali A, Lopez de Cerain Salsamendi A, et al. Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: the challenge ahead. Carcinogenesis. 2015 Jun;36 Suppl 1:S254-96. PubMed PMID: 26106142. Pubmed Central PMCID: PMC4480130. Epub 2015/06/25. eng.
91. Risch A, Wallace DM, Bathers S, Sim E. Slow N-acetylation genotype is a susceptibility factor in occupational and smoking related bladder cancer. Hum Mol Genet. 1995 Feb;4(2):231-6. PubMed PMID: 7757072. Epub 1995/02/01. eng.
92. Letašiová S, Medveďová A, Šovčíková A, Dušinská M, Volkovová K, Mosoiu C, et al. Bladder cancer, a review of the environmental risk factors. Environ Health. 2012 06/28;11(Suppl 1):S11-S. PubMed PMID: PMC3388449.
93. Perera FP, Weinstein IB. Molecular epidemiology and carcinogen-DNA adduct detection: new approaches to studies of human cancer causation. J Chronic Dis. 1982;35(7):581-600. PubMed PMID: 6282919.
94. Perera FP, Weinstein IB. Molecular epidemiology: recent advances and future directions. Carcinogenesis. 2000 Mar;21(3):517-24. PubMed PMID: 10688872. Epub 2000/02/26. eng.
95. Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev. 2007 Oct;16(10):1954-65. PubMed PMID: 17932342.
96. Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. Nature reviews Cancer. 2005 Nov;5(11):845-56. PubMed PMID: 16239904. Epub 2005/10/22. eng.
97. Mayeux R. Biomarkers: Potential Uses and Limitations. NeuroRx. 2004;1(2):182-8. PubMed PMID: PMC534923.
98. Srivastava S, Gopal-Srivastava R. Biomarkers in cancer screening: a public health perspective. J Nutr. 2002 Aug;132(8 Suppl):2471S-5S. PubMed PMID: 12163714. Epub 2002/08/07. eng.
99. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008 Jan 30;27(2):157-72; discussion 207-12. PubMed PMID: 17569110.
100. Poste G. Bring on the biomarkers. Nature. 2011 Jan 13;469(7329):156-7. PubMed PMID: 21228852. Epub 2011/01/14. eng.
101. Weinstein JN. 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer. Curr Opin Pharm. 2002 8/1/;2(4):361-5.

102. Hammond ME, Taube SE. Issues and barriers to development of clinically useful tumor markers: a development pathway proposal. Semin Oncol. 2002 Jun;29(3):213-21. PubMed PMID: 12063674. Epub 2002/06/14. eng.
103. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics. 2002 Nov;1(11):845-67. PubMed PMID: 12488461. Epub 2002/12/19. eng.
104. Malottki K, Biswas M, Deeks JJ, Riley RD, Craddock C, Johnson P, et al. Stratified medicine in European Medicines Agency licensing: a systematic review of predictive biomarkers. BMJ open. 2014;4(1):e004188. PubMed PMID: 24468721. Pubmed Central PMCID: 3913033. Epub 2014/01/29. eng.
105. Beltran H, Rubin MA. New strategies in prostate cancer: translating genomics into the clinic. Clin Cancer Res. 2013 Feb 1;19(3):517-23. PubMed PMID: 23248095. Epub 2012/12/19. eng.
106. Wild CP. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiology Biomarkers and Prevention. 2005;14(8):1847-50.
107. Rappaport SM. Implications of the exposome for exposure science. Journal Of Exposure Science And Environmental Epidemiology. 2010 11/17/online;21(1):5. PubMed PMID: 21081972.
108. Wild CP, Scalbert A, Herceg Z. Measuring the exposome: a powerful basis for evaluating environmental exposures and cancer risk. Environ Mol Mutagen. 2013 Aug;54(7):480-99. PubMed PMID: 23681765. Epub 2013/05/18. eng.
109. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012 Feb;41(1):24-32. PubMed PMID: 22296988. Epub 2012/02/03. Eng.
110. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. PLoS ONE. 2010;5(5):e10746.
111. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, et al. The human early-life exposome (HELIX): project rationale and design. Environ Health Perspect. 2014 Jun;122(6):535-44. PubMed PMID: 24610234. Pubmed Central PMCID: PMC4048258. Epub 2014/03/13. eng.
112. Holmes E, Loo RL, Stamler J, Bictash M, Yap IKS, Chan Q, et al. Human metabolic phenotype diversity and its association with diet and blood pressure. Nature. 2008 04/20/online;453:396.
113. Smith MT, de la Rosa R, Daniels SI. Using exposomics to assess cumulative risks and promote health. Environ Mol Mutagen. 2015 Dec;56(9):715-23. PubMed PMID: 26475350. Pubmed Central PMCID: PMC4636923. Epub 2015/10/18. eng.
114. Stingone JA, Buck Louis GM, Nakayama SF, Vermeulen RC, Kwok RK, Cui Y, et al. Toward Greater Implementation of the Exposome Research Paradigm within Environmental Epidemiology. Annu Rev Public Health. 2017 Mar 20;38(1):315-27. PubMed PMID: 28125387.
115. Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, et al. The exposome in practice: Design of the EXPOsOMICS project. Int J Hyg Environ Health. 2016 Aug 19. PubMed PMID: 27576363. Epub 2016/09/01. eng.
116. Goldfarb DS. The exposome for kidney stones. Urolithiasis. 2016;44(1):3-7.
117. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The blood exposome and its role in discovering causes of disease. Environ Health Perspect.

2014 Aug;122(8):769-74. PubMed PMID: 24659601. Pubmed Central PMCID: 4123034.
118. Wild N, Andres H, Rollinger W, Krause F, Dilba P, Tacke M, et al. A combination of serum markers for the early detection of colorectal cancer. Clin Cancer Res. 2010 Dec 15;16(24):6111-21. PubMed PMID: 20798228. Epub 2010/08/28. eng.
119. Nicholson G, Rantalainen M, Maher AD, Li JV, Malmodin D, Ahmadi KR, et al. Human metabolic profiles are stably controlled by genetic and environmental variation. Mol Syst Biol. 2011 Aug 30;7:525. PubMed PMID: 21878913. Pubmed Central PMCID: PMC3202796. Epub 2011/09/01. Eng.
120. Bellazzi R, Diomidous M, Sarkar IN, Takabayashi K, Ziegler A, McCray AT. Data analysis and data mining: current issues in biomedical informatics. Methods Inf Med. 2011;50(6):536-44. PubMed PMID: 22146916. Pubmed Central PMCID: 3233983. Epub 2011/12/08. eng.
121. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics. 2015 01/13/online;16:85.
122. Crick F. Central Dogma of Molecular Biology. Nature. 1970 08/08/online;227(5258):561. PubMed PMID: 4913914.
123. Fodeh SJ, Brandt C, Luong TB, Haddad A, Schultz M, Murphy T, et al. Complementary ensemble clustering of biomedical data. J Biomed Inform. 2013 Jun;46(3):436-43. PubMed PMID: 23454721. Pubmed Central PMCID: PMC4007219. Epub 2013/03/05. eng.
124. Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, et al. Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers. Environ Mol Mutag. 2013;54(7):542-57.
125. Balding DJ. A tutorial on statistical methods for population association studies. Nature Reviews Genetics. 2006 10/01/online;7(10):781. PubMed PMID: 16983374.
126. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet. 2010;86(1):6-22. PubMed PMID: PMC2801749.
127. Schneider A, Hommel G, Blettner M. Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. Deutsches Ärzteblatt International. 2010 11/05
05/11/received
07/14/accepted;107(44):776-82. PubMed PMID: PMC2992018. Pubmed Central PMCID: 2992018.
128. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annu Rev Public Health. 1999;20:145-57. PubMed PMID: 10352854. Epub 1999/06/03. eng.
129. Kim H-Y. Analysis of variance (ANOVA) comparing means of more than two groups. Restorative Dentistry & Endodontics. 2014 01/20;39(1):74-7. PubMed PMID: PMC3916511. Pubmed Central PMCID: 3916511.
130. McDonald JH. Handbook of biological statistics: Sparky House Publishing Baltimore, MD; 2009.

131. McHugh ML. Multiple comparison analysis testing in ANOVA. Biochemia medica. 2011;21(3):203-9. PubMed PMID: 22420233. Epub 2011/01/01. eng.
132. Kao LS, Green CE. Analysis of Variance: Is There a Difference in Means and What Does It Mean? The Journal of surgical research. 2008 10/22;144(1):158-70. PubMed PMID: PMC2405942.
133. Shaffer JP. Multiple Hypothesis Testing. Annu Rev Psychol. 1995;46(1):561-84. PubMed PMID: WOS:A1995QF07700021. English.
134. Oberg AL, Mahoney DW. Linear mixed effects models. Methods Mol Biol. 2007;404:213-34. PubMed PMID: 18450052. Epub 2008/05/03. eng.
135. Dean CB, Nielsen JD. Generalized linear mixed models: a review and some extensions. Lifetime Data Anal. 2007 Dec;13(4):497-512. PubMed PMID: 18000755. Epub 2007/11/15. eng.
136. Mukaka MM. A guide to appropriate use of Correlation coefficient in medical research. Malawi Medical Journal : The Journal of Medical Association of Malawi. 2012;24(3):69-71. PubMed PMID: PMC3576830.
137. Zhang L, Pei Y-F, Li J, Papasian CJ, Deng H-W. Univariate/Multivariate Genome-Wide Association Scans Using Data from Families and Unrelated Samples. PLoS ONE. 2009 08/04
04/12/received
06/30/accepted;4(8):e6502. PubMed PMID: PMC2715864.
138. Li L. Dimension reduction for high-dimensional data. Methods Mol Biol. 2010;620:417-34. PubMed PMID: 20652514. Epub 2010/07/24. eng.
139. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901;2(11):559-72.
140. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics Intellig Lab Syst. 1987;2(1-3):37-52.
141. Barber DC, Howlett PJ, Smart RC. Principal component analysis in medical research. Journal of Applied Statistics. 1975 1975/01/01;2(1):39-43.
142. Ringner M. What is principal component analysis? Nat Biotechnol. 2008 Mar;26(3):303-4. PubMed PMID: 18327243. Epub 2008/03/11. eng.
143. Suhr DD. Principal component analysis vs. exploratory factor analysis. SUGI 30 proceedings. 2005;203:230.
144. Cattell R. The scientific use of factor analysis in behavioral and life sciences: Springer Science & Business Media; 2012.
145. Fruchter B. Introduction to factor analysis. 1954.
146. Manhenke C, Orn S, von Haehling S, Wollert KC, Ueland T, Aukrust P, et al. Clustering of 37 circulating biomarkers by exploratory factor analysis in patients following complicated acute myocardial infarction. Int J Cardiol. 2013 Jul 1;166(3):729-35. PubMed PMID: 22197217. Epub 2011/12/27. eng.
147. Pladevall M, Singal B, Williams LK, Brotons C, Guyer H, Sadurni J, et al. A single factor underlies the metabolic syndrome: a confirmatory factor analysis. Diabetes Care. 2006 Jan;29(1):113-22. PubMed PMID: 16373906. Epub 2005/12/24. eng.
148. Woolston A, Tu YK, Baxter PD, Gilthorpe MS. A comparison of different approaches to unravel the latent structure within metabolic syndrome. PLoS ONE.

2012;7(4):e34410. PubMed PMID: 22485169. Pubmed Central PMCID: PMC3317545. Epub 2012/04/10. eng.
149. Wold S, Ruhe A, Wold H, Dunn I, WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing. 1984;5(3):735-43.
150. Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics. 2002 Sep;18(9):1216-26. PubMed PMID: 12217913. Epub 2002/09/10. eng.
151. Thayer JF. A hybrid approach to clustering biomedical data. Biomed Sci Instrum. 1996;32:39-46. PubMed PMID: 8672688. Epub 1996/01/01. eng.
152. Doostparast Torshizi A, Fazel Zarandi MH. A new cluster validity measure based on general type-2 fuzzy sets: Application in gene expression data clustering. Knowledge-Based Systems. 2014 7//;64(0):81-93.
153. Belacel N, Wang Q, Cuperlovic-Culf M. Clustering methods for microarray gene expression data. Omics : a journal of integrative biology. 2006 Winter;10(4):507-31. PubMed PMID: 17233561. Epub 2007/01/20. eng.
154. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001;17(10):977-87.
155. Kostering L, McKinlay A, Stahl C, Kaller CP. Differential patterns of planning impairments in Parkinson's disease and sub-clinical signs of dementia? A latent-class model-based approach. PLoS ONE. 2012;7(6):e38855. PubMed PMID: 22715417. Pubmed Central PMCID: PMC3371002. Epub 2012/06/21. eng.
156. McCutcheon AL. Latent class analysis: Sage; 1987.
157. Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal. 1998;41(8):578-88.
158. Everitt B, Landau S, Leese M, Stahl D, Everitt B, Landau S, et al. Cluster analysis. Chichester: Wiley; 2011.
159. Vermunt JK, Magidson J. Latent class cluster analysis. Applied latent class analysis. 2002:89-106.
160. Haughton D, Legrand P, Woolford S. Review of three latent class cluster analysis packages: Latent Gold, poLCA, and MCLUST. The American Statistician. 2009;63(1).
161. Wood PK. J. A. Hagenaars and A. L. McCutcheon, Applied Latent Class Analysis, Kluwer, Dordrecht, 2002, pp. 476. Journal of Classification. 2008 2008/06/01;25(1):143-5. English.
162. Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A SAS Procedure for Latent Class Analysis. Structural equation modeling : a multidisciplinary journal. 2007;14(4):671-94. PubMed PMID: 19953201. Pubmed Central PMCID: PMC2785099. Epub 2007/01/01. Eng.
163. Linzer DA, Lewis JB. poLCA: An R package for polytomous variable latent class analysis.
164. Fraley C, Raftery AE. MCLUST: Software for model-based cluster analysis. Journal of Classification. 1999;16(2):297-306.
165. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association. 2002;97(458):611-31.

166. Shalabi A, Inoue M, Watkins J, De Rinaldis E, Coolen AC. Bayesian clinical classification from high-dimensional data: Signatures versus variability. Stat Methods Med Res. 2016 Mar 16. PubMed PMID: 26984907. Epub 2016/03/18. eng.
167. Walldius G, Malmstrom H, Jungner I, de Faire U, Lambe M, Van Hemelrijck M, et al. The AMORIS cohort. Int J Epidemiol. 2017 Feb 02;46(4):1103-i. PubMed PMID: 28158674. Epub 2017/02/06. eng.
168. Walldius G MH, Jungner I, de Faire U, Lambe M, Van Hemelrijck M, Hammar N. The AMORIS cohort – more than 800,000 subjects with information on biomarkers followed for more than 2- years in Swedish national health and quality of care registers. Int J Epidemiology. 2016;In press.
169. Van Hemelrijck M, Garmo H, Binda E, Hayday A, Karagiannis SN, Hammar N, et al. Immunoglobulin E and cancer: a meta-analysis and a large Swedish cohort study. Cancer Causes Control. 2010 Oct;21(10):1657-67. PubMed PMID: 20533084. Epub 2010/06/10. eng.
170. Van Hemelrijck M, Holmberg L, Garmo H, Hammar N, Walldius G, Binda E, et al. Association between levels of C-reactive protein and leukocytes and cancer: three repeated measurements in the Swedish AMORIS study. Cancer Epidemiol Biomarkers Prev. 2011 Mar;20(3):428-37. PubMed PMID: 21297038. Pubmed Central PMCID: 3078551. Epub 2011/02/08. eng.
171. Van Hemelrijck M, Garmo H, Holmberg L, Walldius G, Jungner I, Hammar N, et al. Prostate cancer risk in the Swedish AMORIS study: the interplay among triglycerides, total cholesterol, and glucose. Cancer. 2011 May 15;117(10):2086-95. PubMed PMID: 21523720. Epub 2011/04/28. eng.
172. Van Hemelrijck M, Jassem W, Walldius G, Fentiman IS, Hammar N, Lambe M, et al. Gamma-glutamyltransferase and risk of cancer in a cohort of 545,460 persons - the Swedish AMORIS study. Eur J Cancer. 2011 Sep;47(13):2033-41. PubMed PMID: 21486691. Epub 2011/04/14. eng.
173. Van Hemelrijck M, Walldius G, Jungner I, Hammar N, Garmo H, Binda E, et al. Low levels of apolipoprotein A-I and HDL are associated with risk of prostate cancer in the Swedish AMORIS study. Cancer Causes Control. 2011 Jul;22(7):1011-9. PubMed PMID: 21562751. Epub 2011/05/13. eng.
174. Melvin JC, Seth D, Holmberg L, Garmo H, Hammar N, Jungner I, et al. Lipid profiles and risk of breast and ovarian cancer in the Swedish AMORIS study. Cancer Epidemiol Biomarkers Prev. 2012 Aug;21(8):1381-4. PubMed PMID: 22593241. Epub 2012/05/18. eng.
175. Seth D, Garmo H, Wigertz A, Holmberg L, Hammar N, Jungner I, et al. Lipid profiles and the risk of endometrial cancer in the Swedish AMORIS study. International journal of molecular epidemiology and genetics. 2012;3(2):122-33. PubMed PMID: 22724049. Pubmed Central PMCID: PMC3376923. Epub 2012/06/23. eng.
176. Van Hemelrijck M, Garmo H, Hammar N, Jungner I, Walldius G, Lambe M, et al. The interplay between lipid profiles, glucose, BMI and risk of kidney cancer in the Swedish AMORIS study. Int J Cancer. 2012 May 1;130(9):2118-28. PubMed PMID: 21630265. Epub 2011/06/02. eng.
177. Van Hemelrijck M, Hermans R, Michaelsson K, Melvin J, Garmo H, Hammar N, et al. Serum calcium and incident and fatal prostate cancer in the Swedish

AMORIS study. Cancer Causes Control. 2012 Aug;23(8):1349-58. PubMed PMID: 22710746. Epub 2012/06/20. eng.
178. Wulaningsih W, Garmo H, Holmberg L, Hammar N, Jungner I, Walldius G, et al. Serum Lipids and the Risk of Gastrointestinal Malignancies in the Swedish AMORIS Study. Journal of cancer epidemiology. 2012;2012:792034. PubMed PMID: 22969802. Pubmed Central PMCID: 3437288.
179. Gaur A, Collins H, Wulaningsih W, Holmberg L, Garmo H, Hammar N, et al. Iron metabolism and risk of cancer in the Swedish AMORIS study. Cancer Causes Control. 2013 Jul;24(7):1393-402. PubMed PMID: 23649231. Pubmed Central PMCID: 3675271. Epub 2013/05/08. eng.
180. Wulaningsih W, Michaelsson K, Garmo H, Hammar N, Jungner I, Walldius G, et al. Inorganic phosphate and the risk of cancer in the Swedish AMORIS study. BMC Cancer. 2013;13:257. PubMed PMID: 23706176. Pubmed Central PMCID: PMC3664604. Epub 2013/05/28. eng.
181. Wulaningsih W, Michaelsson K, Garmo H, Hammar N, Jungner I, Walldius G, et al. Serum calcium and risk of gastrointestinal cancer in the Swedish AMORIS study. BMC Public Health. 2013 Jul 17;13(1):663. PubMed PMID: 23866097. Pubmed Central PMCID: 3729677. Epub 2013/07/20. Eng.
182. Wulaningsih W, Holmberg L, Garmo H, Zethelius B, Wigertz A, Carroll P, et al. Serum glucose and fructosamine in relation to risk of cancer. PLoS ONE. 2013;8(1):e54944. PubMed PMID: 23372798. Pubmed Central PMCID: PMC3556075. Epub 2013/02/02. eng.
183. Arthur R, Moller H, Garmo H, Holmberg L, Stattin P, Malmstrom H, et al. Association between baseline serum glucose, triglycerides and total cholesterol, and prostate cancer risk categories. Cancer medicine. 2016 Jun;5(6):1307-18. PubMed PMID: 26923095. Pubmed Central PMCID: 4924389. Epub 2016/03/01. eng.
184. Ghoshal A, Garmo H, Arthur R, Hammar N, Jungner I, Malmström H, et al. Serum biomarkers to predict risk of testicular and penile cancer in AMORIS. ecancermedicalscience. 2017;11:762. PubMed PMID: 28900475. Pubmed Central PMCID: 5574660.
185. Walldius G, Jungner I, Kolar W, Holme I, Steiner E. High cholesterol and triglyceride values in Swedish males and females: increased risk of fatal myocardial infarction. First report from the AMORIS (Apolipoprotein related MOrtality RISk) study. Blood Press Suppl. 1992;4:35-42. PubMed PMID: 1345333. Epub 1992/01/01. eng.
186. Jungner I, Marcovina SM, Walldius G, Holme I, Kolar W, Steiner E. Apolipoprotein B and A-I values in 147576 Swedish males and females, standardized according to the World Health Organization-International Federation of Clinical Chemistry First International Reference Materials. Clin Chem. 1998 Aug;44(8 Pt 1):1641-9. PubMed PMID: 9702950. Epub 1998/08/14. eng.
187. Walldius G, Jungner I, Holme I, Aastveit AH, Kolar W, Steiner E. High apolipoprotein B, low apolipoprotein A-I, and improvement in the prediction of fatal myocardial infarction (AMORIS study): a prospective study. Lancet. 2001 Dec 15;358(9298):2026-33. PubMed PMID: 11755609. Epub 2002/01/05. eng.
188. Holme I, Aastveit AH, Hammar N, Jungner I, Walldius G. Lipoprotein components and risk of congestive heart failure in 84,740 men and women in the

Apolipoprotein MOrtality RISk study (AMORIS). European journal of heart failure. 2009 Nov;11(11):1036-42. PubMed PMID: 19801574. Epub 2009/10/06. eng.
189. Berg JM, Tymoczko JL, Stryer L. Biochemistry: Macmillan; 2008.
190. Stryer L. Fatty acid metabolism. Biochemistry (Mosc). 1995;3:469-93.
191. Harris MI, Flegal KM, Cowie CC, Eberhardt MS, Goldstein DE, Little RR, et al. Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in US adults: the Third National Health and Nutrition Examination Survey, 1988–1994. Diabetes Care. 1998;21(4):518-24.
192. Bartol T. Comparison of blood glucose, HbA1c, and fructosamine. Feb; 2009.
193. Assmann G, Schulte H. Relation of high-density lipoprotein cholesterol and triglycerides to incidence of atherosclerotic coronary artery disease (the PROCAM experience). The American journal of cardiology. 1992;70(7):733-7.
194. Kuyl JM, Mendelsohn D. Observed relationship between ratios HDL-cholesterol/total cholesterol and apolipoprotein A1/apolipoprotein B. Clin Biochem. 1992 Oct;25(5):313-6. PubMed PMID: 1490290. Epub 1992/10/01. eng.
195. Click Jr JH. Serum lactate dehydrogenase isoenzyme and total lactate dehydrogenase values in health and disease, and clinical evaluation of these tests by means of discriminant analysis. Am J Clin Pathol. 1969;52(3):320-8.
196. Neidle S. Cancer drug design and discovery: Academic Press; 2011.
197. Brauer RW. Liver circulation and function. Physiol Rev. 1963;43(1):115-214.
198. Kim WR, Flamm SL, Di Bisceglie AM, Bodenheimer HC. Serum activity of alanine aminotransferase (ALT) as an indicator of health and disease. Hepatology. 2008;47(4):1363-70.
199. Karmen A, Wróblewski F, LaDue JS. Transaminase activity in human blood. J Clin Invest. 1955;34(1):126.
200. Lum G, Gambino SR. Serum gamma-glutamyl transpeptidase activity as an indicator of disease of liver, pancreas, or bone. Clin Chem. 1972;18(4):358-62.
201. Mason JE, Starke RD, Van Kirk JE. Gamma-glutamyl transferase: a novel cardiovascular risk biomarker. Preventive cardiology. 2010 Winter;13(1):36-41. PubMed PMID: 20021625. Epub 2009/12/22. eng.
202. Millan J. Alkaline phosphatases structure, substrate specificity and functional relatedness to other members of a large superfamily of enzymes Purinergic Signal 2006 2. N.
203. Vize PD, Woolf AS, Bard JB. The kidney: from normal development to congenital disease: Academic Press; 2003.
204. Cockcroft DW, Gault H. Prediction of creatinine clearance from serum creatinine. Nephron. 1976;16(1):31-41.
205. Ferguson JD, Galligan DT, Blanchard T, Reeves M. Serum urea nitrogen and conception rate: the usefulness of test information. J Dairy Sci. 1993;76(12):3742-6.
206. McCrudden FH. Uric Acid: BiblioBazaar, LLC; 2008.
207. Johnson RJ, Kang D-H, Feig D, Kivlighn S, Kanellis J, Watanabe S, et al. Is there a pathogenetic role for uric acid in hypertension and cardiovascular and renal disease? Hypertension. 2003;41(6):1183-90.
208. Jee SH, Lee SY, Kim MT. Serum uric acid and risk of death from cancer, cardiovascular disease or all causes in men. European Journal of Cardiovascular Prevention & Rehabilitation. 2004;11(3):185-91.

209. Buchmann K. Evolution of innate immunity: clues from invertebrates via fish to mammals. Frontiers in immunology. 2014;5.
210. Peters Jr T. All about albumin: biochemistry, genetics, and medical applications: Academic press; 1995.
211. Arroyo V, García-Martinez R, Salvatella X. Human serum albumin, systemic inflammation, and cirrhosis. J Hepatol. 2014 2014/08/01/;61(2):396-407.
212. Pepys MB, Hirschfield GM. C-reactive protein: a critical update. J Clin Invest. 2003;111(12):1805.
213. Ridker PM, Buring JE, Cook NR, Rifai N. C-reactive protein, the metabolic syndrome, and risk of incident cardiovascular events. Circulation. 2003;107(3):391-7.
214. Langlois MR, Delanghe JR. Biological and clinical significance of haptoglobin polymorphism in humans. Clin Chem. 1996;42(10):1589-600.
215. Kelly C. Can excess iron increase the risk for coronary heart disease and cancer? Nutrition Bulletin. 2002;27(3):165-79. English.
216. Prutki M, Poljak-Blazi M, Jakopovic M, Tomas D, Stipancic I, Zarkovic N. Altered iron metabolism, transferrin receptor 1 and ferritin in patients with colon cancer. Cancer Lett. 2006 Jul 18;238(2):188-96. PubMed PMID: 16111806. Epub 2005/08/23. eng.
217. Kabat GC, Rohan TE. Does excess iron play a role in breast carcinogenesis? An unresolved hypothesis. Cancer Causes Control. 2007 Dec;18(10):1047-53. PubMed PMID: 17823849. Epub 2007/09/08. eng.
218. Mascitelli L, Pezzetta F, Goldstein MR. Diabetes, cancer and iron. Diabetologia. 2010 Sep;53(9):2071-2. PubMed PMID: 20567802. Epub 2010/06/23. eng.
219. Munoz M, Garcia-Erce JA, Remacha AF. Disorders of iron metabolism. Part II: iron deficiency and iron overload. J Clin Pathol. 2011 Apr;64(4):287-96. PubMed PMID: 21177268. English.
220. Datz C, Felder TK, Niederseer D, Aigner E. Iron homeostasis in the metabolic syndrome. Eur J Clin Invest. 2013 Feb;43(2):215-24. PubMed PMID: 23289518. English.
221. Wang J, Pantopoulos K. Regulation of cellular iron metabolism. Biochem J. 2011 Mar 15;434(3):365-81. PubMed PMID: 21348856. Pubmed Central PMCID: PMC3048577. Epub 2011/02/26. eng.
222. Waldvogel-Abramowski S, Waeber G, Gassner C, Buser A, Frey BM, Favrat B, et al. Physiology of iron metabolism. Transfusion medicine and hemotherapy : offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie. 2014 Jun;41(3):213-21. PubMed PMID: 25053935. Pubmed Central PMCID: 4086762.
223. Yamanishi H, Iyama S, Yamaguchi Y, Kanakura Y, Iwatani Y. Total iron-binding capacity calculated from serum transferrin concentration or serum iron concentration and unsaturated iron-binding capacity. Clin Chem. 2003;49(1):175-8.
224. Shibata H. Cancer and electrolytes imbalance. Gan to kagaku ryoho Cancer & chemotherapy. 2010 Jun;37(6):1006-10. PubMed PMID: 20567101.
225. Shanahan CM, Crouthamel MH, Kapustin A, Giachelli CM. Arterial calcification in chronic kidney disease: key roles for calcium and phosphate. Circul Res. 2011;109(6):697-711.

226. Moe SM. Disorders involving calcium, phosphorus, and magnesium. Primary Care: Clinics in Office Practice. 2008 Jun;35(2):215-37. PubMed PMID: 18486714. Pubmed Central PMCID: 2486454.
227. Danowski TS. Newer concepts of the role of potassium in disease. The American Journal of Medicine. 1949 1949/10/01/;7(4):525-31. PubMed PMID: 18140550.
228. Schwartz DA. The importance of gene-environment interactions and exposure assessment in understanding human diseases. Journal of Exposure Science and Environmental Epidemiology. 2006;16(6):474-.
229. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiology and Prevention Biomarkers. 2006;15(6):1159-69.
230. Sandhu MS, White IR, McPherson K. Systematic review of the prospective cohort studies on meat consumption and colorectal cancer risk. Cancer Epidemiology and Prevention Biomarkers. 2001;10(5):439-46.
231. Catalona WJ, Richie JP, Ahmann FR, M'Liss AH, Scardino PT, Flanigan RC, et al. Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men. The Journal of urology. 1994;151(5):1283-90.
232. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994;266(5182):66-71.
233. Peters A, Hoek G, Katsouyanni K. Understanding the link between environmental exposures and health: does the exposome promise too much? J Epidemiol Community Health. 2012 Feb;66(2):103-5. PubMed PMID: 22080817. Epub 2011/11/15. eng.
234. Brunekreef B. Exposure science, the exposome, and public health. Environ Mol Mutagen. 2013 Aug;54(7):596-8. PubMed PMID: 23444186. Epub 2013/02/28. eng.
235. Nakamura J, Mutlu E, Sharma V, Collins L, Bodnar W, Yu R, et al. The endogenous exposome. DNA Repair (Amst). 2014 Jul;19:3-13. PubMed PMID: 24767943. Pubmed Central PMCID: PMC4097170. Epub 2014/04/29. eng.
236. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. Toxicol Sci. 2014 Jan;137(1):1-2. PubMed PMID: 24213143. Pubmed Central PMCID: PMC3871934. Epub 2013/11/12. eng.
237. Siroux V, Agier L, Slama R. The exposome concept: a challenge and a potential driver for environmental health research. European respiratory review : an official journal of the European Respiratory Society. 2016 Jun;25(140):124-9. PubMed PMID: 27246588. Epub 2016/06/02. eng.
238. Cui Y, Balshaw DM, Kwok RK, Thompson CL, Collman GW, Birnbaum LS. The Exposome: Embracing the Complexity for Discovery in Environmental Health. Environ Health Perspect. 2016 Aug 01;124(8):A137-40. PubMed PMID: 27479988. Pubmed Central PMCID: PMC4977033. Epub 2016/08/02. eng.
239. Arjun K. Manrai YC, 2 Pierre R. Bushel,2, Molly Hall SK, 4 Carolyn J. Mattingly,5, Marylyn Ritchie, 6 Charles Schmitt,7, Denis A. Sarigiannis DCT, 8, David Wishart DMB, 2, Patel aCJ. Informatics and Data Analytics to Support Exposome-

Based Discovery for Public Health. Annu Rev Public Health. 2017;38(1):null. PubMed PMID: 28068484.
240. Louis GMB, Smarr MM, Patel CJ. The exposome research paradigm: an opportunity to understand the environmental basis for human health and disease. Current environmental health reports. 2017;4(1):89-98.
241. Holme I, Aastveit AH, Hammar N, Jungner I, Walldius G. Inflammatory markers, lipoprotein components and risk of major cardiovascular events in 65,005 men and women in the Apolipoprotein MOrtality RISk study (AMORIS). Atherosclerosis. 2010 Nov;213(1):299-305. PubMed PMID: 20843515.
242. McLachlan GJ. Cluster analysis and related techniques in medical research. Stat Methods Med Res. 1992;1(1):27-48. PubMed PMID: 1341650. Epub 1992/01/01. eng.
243. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016;374(2065).
244. Chadeau-Hyam M, Athersuch TJ, Keun HC, De Iorio M, Ebbels TMD, Jenab M, et al. Meeting-in-the-middle using metabolic profiling - a strategy for the identification of intermediate biomarkers in cohort studies. Biomarkers. 2011 Feb;16(1):83-8. PubMed PMID: WOS:000287451400011. English.
245. Wei T, Simko V. Package 'corrplot'. R package version 0.77. 2016.
246. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006;1695(5):1-9.
247. Wehrens R, Mevik B-H. The pls package: principal component and partial least squares regression in R. 2007.
248. Studio R. RStudio: integrated development environment for R. RStudio Inc, Boston, Massachusetts. 2012.
249. Cody RP, Smith JK. Applied statistics and the SAS programming language. New York: North-Holland; 1985 1985.
250. Gkouvatsos K, Papanikolaou G, Pantopoulos K. Regulation of iron transport and the role of transferrin. Biochimica et Biophysica Acta (BBA)-General Subjects. 2012;1820(3):188-202.
251. Xie J-H, Liu Q, Yang Y, Liu Z-L, Hu S-H, Zhou X-R, et al. Correlation of liver enzymes with diabetes and pre-diabetes in middle-aged rural population in China. Journal of Huazhong University of Science and Technology [Medical Sciences]. 2016;36(1):53-8.
252. Nissen NI, Ranløv P, Weis-Fogh J. Evaluation of four different serum enzymes in the diagnosis of acute myocardial infarction. Br Heart J. 1965;27(4):520.
253. Mathey D, Biefield W, Hanrath P, Effert S. Attempt to quantitate relation between cardiac function and infarct size in acute myocardial infarction. Br Heart J. 1974;36(3):271-9. PubMed PMID: PMC458829.
254. Konttinen A, Halonen P. Cardiologia (Basel) 43: 56, 1963. CrossRef| PubMed| CAS.
255. Ridefelt P, Helmersson-Karlqvist J. Albumin adjustment of total calcium does not improve the estimation of calcium status. Scand J Clin Lab Invest. 2017:1-6.
256. Lee J-M, Kim HC, Cho HM, Oh SM, Choi DP, Suh I. Association between serum uric acid level and metabolic syndrome. Journal of Preventive Medicine and Public Health. 2012;45(3):181.

257. Sirota JC, McFann K, Targher G, Johnson RJ, Chonchol M, Jalal DI. Elevated serum uric acid levels are associated with non-alcoholic fatty liver disease independently of metabolic syndrome features in the United States: Liver ultrasound data from the National Health and Nutrition Examination Survey. Metabolism. 2013;62(3):392-9.
258. Ferraresi M, Pia A, Guzzo G, Vigotti FN, Mongilardi E, Nazha M, et al. Calcium-phosphate and parathyroid intradialytic profiles: A potential aid for tailoring the dialysate calcium content of patients on different hemodialysis schedules. Hemodialysis International. 2015;19(4):572-82.
259. Pupim LB, Martin CJ, Ikizler TA. Chapter 10 - Assessment of Protein and Energy Nutritional Status A2 - Kopple, Joel D. In: Massry SG, Kalantar-Zadeh K, editors. Nutritional Management of Renal Disease (Third Edition): Academic Press; 2013. p. 137-58.
260. Hardy OT, Czech MP, Corvera S. What causes the insulin resistance underlying obesity? Current opinion in endocrinology, diabetes, and obesity. 2012;19(2):81.
261. Herman-Edelstein M, Scherzer P, Tobar A, Levi M, Gafter U. Altered renal lipid metabolism and renal lipid accumulation in human diabetic nephropathy. J Lipid Res. 2014;55(3):561-72.
262. Braveman PA, Cubbin C, Egerter S, Williams DR, Pamuk E. Socioeconomic disparities in health in the United States: what the patterns tell us. Am J Public Health. 2010;100(S1):S186-S96.
263. Cutler DM, Lleras-Muney A. Understanding differences in health behaviors by education. J Health Econ. 2010;29(1):1-28.
264. Gordon-Larsen P, Nelson MC, Page P, Popkin BM. Inequality in the built environment underlies key health disparities in physical activity and obesity. Pediatrics. 2006;117(2):417-24.
265. Stringhini S, Sabia S, Shipley M, Brunner E, Nabi H, Kivimaki M, et al. Association of socioeconomic position with health behaviors and mortality. JAMA. 2010;303(12):1159-66.
266. Vasikaran S, Eastell R, Bruyère O, Foldes A, Garnero P, Griesmacher A, et al. Markers of bone turnover for the prediction of fracture risk and monitoring of osteoporosis treatment: a need for international reference standards. Osteoporosis Int. 2011;22(2):391-420.
267. Murphy E. Estrogen signaling and cardiovascular disease. Circul Res. 2011;109(6):687-96.
268. Rachner TD, Khosla S, Hofbauer LC. Osteoporosis: now and the future. The Lancet. 2011;377(9773):1276-87.
269. Chajès V, Thiébaut AC, Rotival M, Gauthier E, Maillard V, Boutron-Ruault M-C, et al. Association between serum trans-monounsaturated fatty acids and breast cancer risk in the E3N-EPIC Study. Am J Epidemiol. 2008;167(11):1312-20.
270. Chajès V, Jenab M, Romieu I, Ferrari P, Dahm CC, Overvad K, et al. Plasma phospholipid fatty acid concentrations and risk of gastric adenocarcinomas in the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST). The American Journal of Clinical Nutrition. 2011 November 1, 2011;94(5):1304-13.

271. Bünger S, Haug U, Kelly M, Posorski N, Klempt-Giessing K, Cartwright A, et al. A novel multiplex-protein array for serum diagnostics of colon cancer: a case–control study. BMC Cancer. 2012 09/07
05/11/received
08/31/accepted;12:393-. PubMed PMID: PMC3502594.
272. Assi N, Fages A, Vineis P, Chadeau-Hyam M, Stepien M, Duarte-Salles T, et al. A statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study. Mutagenesis. 2015 Nov;30(6):743-53. PubMed PMID: 26130468. Epub 2015/07/02. eng.
273. Blair RH, Trichler DL, Gaille DP. Mathematical and statistical modeling in cancer systems biology. Front Physiol. 2012 06/28
03/30/received
06/05/accepted;3:227. PubMed PMID: 22754537. Pubmed Central PMCID: 3385354.
274. Jones DP, Park Y, Ziegler TR. Nutritional metabolomics: Progress in addressing complexity in diet and health. Annu Rev Nutr. 2012 Aug 21;32:183-202. PubMed PMID: 22540256. Pubmed Central PMCID: PMC4031100. Epub 2012/05/01. eng.
275. Shah SH, Sun JL, Stevens RD, Bain JR, Muehlbauer MJ, Pieper KS, et al. Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. Am Heart J. 2012 May;163(5):844-50 e1. PubMed PMID: 22607863. Epub 2012/05/23. Eng.
276. Lacey RJ, Strauss VY, Rathod T, Belcher J, Croft PR, Natvig B, et al. Clustering of pain and its associations with health in people aged 50 years and older: cross-sectional results from the North Staffordshire Osteoarthritis Project. BMJ open. 2015 November 1, 2015;5(11).
277. Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol. 2011 Mar 15;173(6):676-82. PubMed PMID: 21330339.
278. Stone NJ, Robinson J, Lichtenstein AH, Merz CNB, Blum CB, Eckel RH, et al. 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2013 November 12, 2013.
279. Anderson TJ, Grégoire J, Hegele RA, Couture P, Mancini GBJ, McPherson R, et al. 2012 Update of the Canadian Cardiovascular Society Guidelines for the Diagnosis and Treatment of Dyslipidemia for the Prevention of Cardiovascular Disease in the Adult. The Canadian journal of cardiology. 2013;29(2):151-67.
280. Grundy SM, Brewer HB, Cleeman JI, Smith SC, Lenfant C, Participants ftC. Definition of Metabolic Syndrome: Report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on Scientific Issues Related to Definition. Circulation. 2004 January 27, 2004;109(3):433-8.
281. Magidson J, Vermunt JK. Comparing latent class factor analysis with the traditional approach in data mining. Statistical Data Mining and Knowledge Discovery. 2004:373-83. PubMed PMID: WOS:000189408200022. English.

282. Van Hemelrijck M, Harari D, Garmo H, Hammar N, Walldius G, Lambe M, et al. Biomarker-based score to predict mortality in persons aged 50 years and older: a new approach in the Swedish AMORIS study. International journal of molecular epidemiology and genetics. 2012;3(1):66-76. PubMed PMID: 22493753. Pubmed Central PMCID: 3316450. Epub 2012/04/12. eng.
283. Dobiášová M, Frohlich J. The plasma parameter log (TG/HDL-C) as an atherogenic index: correlation with lipoprotein particle size and esterification rate inapob-lipoprotein-depleted plasma (FERHDL). Clin Biochem. 2001 10//;34(7):583-8.
284. Dobiasova M, Frohlich J, Sedova M, Cheung MC, Brown BG. Cholesterol esterification and atherogenic index of plasma correlate with lipoprotein size and findings on coronary angiography. J Lipid Res. 2011 Mar;52(3):566-71. PubMed PMID: 21224290. Pubmed Central PMCID: 3035693. Epub 2011/01/13. eng.
285. Sniderman AD, Holme I, Aastveit A, Furberg C, Walldius G, Jungner I. Relation of age, the apolipoprotein B/apolipoprotein A-I ratio, and the risk of fatal myocardial infarction and implications for the primary prevention of cardiovascular disease. Am J Cardiol. 2007 Jul 15;100(2):217-21. PubMed PMID: 17631073. Epub 2007/07/17. eng.
286. Tolonen H, Keil U, Ferrario M, Evans A, Project WM. Prevalence, awareness and treatment of hypercholesterolaemia in 32 populations: results from the WHO MONICA Project. Int J Epidemiol. 2005 Feb;34(1):181-92. PubMed PMID: 15333620.
287. Reiner Ž, Catapano AL, De Backer G, Graham I, Taskinen M-R, Wiklund O, et al. ESC/EAS Guidelines for the management of dyslipidaemias. The Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and the European Atherosclerosis Society (EAS). 2011 2011-07-01 00:00:00;32(14):1769-818.
288. Ioannou GN, Boyko EJ, Lee SP. The prevalence and predictors of elevated serum aminotransferase activity in the United States in 1999-2002. Am J Gastroenterol. 2006 Jan;101(1):76-82. PubMed PMID: 16405537. Epub 2006/01/13. eng.
289. Teppala S, Shankar A, Li J, Wong TY, Ducatman A. Association between serum gamma-glutamyltransferase and chronic kidney disease among US adults. Kidney Blood Press Res. 2010;33(1):1-6. PubMed PMID: 20090360. Epub 2010/01/22. eng.
290. Lim JS, Yang JH, Chun BY, Kam S, Jacobs DR, Jr., Lee DH. Is serum gamma-glutamyltransferase inversely associated with serum antioxidants as a marker of oxidative stress? Free Radic Biol Med. 2004 Oct 01;37(7):1018-23. PubMed PMID: 15336318. Epub 2004/09/01. eng.
291. Wessling-Resnick M. Iron Homeostasis and the Inflammatory Response. Annu Rev Nutr. 2010;30:105-22. PubMed PMID: PMC3108097.
292. Dicato M. Anemia in cancer: some pathophysiological aspects. Oncologist. 2003;8(Supplement 1):19-21.
293. Macciò A, Madeddu C, Gramignano G, Mulas C, Tanca L, Cherchi MC, et al. The role of inflammation, iron, and nutritional status in cancer-related anemia: results of a large, prospective, observational study. Haematologica. 2015 06/25/received
09/16/accepted;100(1):124-32. PubMed PMID: PMC4281325.

294. Adamson JW. The anemia of inflammation/malignancy: mechanisms and management. ASH Education Program Book. 2008;2008(1):159-65.
295. Strasak AM, Rapp K, Brant LJ, Hilbe W, Gregory M, Oberaigner W, et al. Association of gamma-glutamyltransferase and risk of cancer incidence in men: a prospective study. Cancer Res. 2008 May 15;68(10):3970-7. PubMed PMID: 18483283. Epub 2008/05/17. eng.
296. Strasak AM, Pfeiffer RM, Klenk J, Hilbe W, Oberaigner W, Gregory M, et al. Prospective study of the association of gamma-glutamyltransferase with cancer incidence in women. Int J Cancer. 2008 Oct 15;123(8):1902-6. PubMed PMID: 18688855. Epub 2008/08/09. eng.
297. Ruhl CE, Everhart JE. Elevated serum alanine aminotransferase and gamma-glutamyltransferase and mortality in the United States population. Gastroenterology. 2009 Feb;136(2):477-85 e11. PubMed PMID: 19100265. Epub 2008/12/23. eng.
298. Koehler EM, Sanna D, Hansen BE, van Rooij FJ, Heeringa J, Hofman A, et al. Serum liver enzymes are associated with all-cause mortality in an elderly population. Liver international : official journal of the International Association for the Study of the Liver. 2014 Feb;34(2):296-304. PubMed PMID: 24219360. Epub 2013/11/14. eng.
299. Kunutsor SK, Apekey TA, Seddoh D, Walley J. Liver enzymes and risk of all-cause mortality in general populations: a systematic review and meta-analysis. Int J Epidemiol. 2014 February 1, 2014;43(1):187-201.
300. De Santis M, Crotti C, Selmi C. Liver abnormalities in connective tissue diseases. Best Practice & Research Clinical Gastroenterology. 2013 2013/08/01/;27(4):543-51.
301. Rose G, Shipley MJ. Plasma lipids and mortality: a source of error. Lancet. 1980 Mar 08;1(8167):523-6. PubMed PMID: 6102243. Epub 1980/03/08. eng.
302. Schupf N, Costa R, Luchsinger J, Tang MX, Lee JH, Mayeux R. Relationship between plasma lipids and all-cause mortality in nondemented elderly. J Am Geriatr Soc. 2005 Feb;53(2):219-26. PubMed PMID: 15673344. Epub 2005/01/28. eng.
303. Akerblom JL, Costa R, Luchsinger JA, Manly JJ, Tang M-X, Lee JH, et al. Relation of plasma lipids to all-cause mortality in Caucasian, African-American and Hispanic elders. Age Ageing. 2008;37(2):207-13. PubMed PMID: PMC2715146.
304. Neaton JD, Blackburn H, Jacobs D, et al. Serum cholesterol level and mortality findings for men screened in the multiple risk factor intervention trial. Arch Intern Med. 1992;152(7):1490-500.
305. Kagan A, McGee DL, Yano K, Rhoads GG, Nomura A. Serum cholesterol and mortality in a Japanese-American population: the Honolulu Heart program. Am J Epidemiol. 1981 Jul;114(1):11-20. PubMed PMID: 7246518. Epub 1981/07/01. eng.
306. Radišauskas R, Kuzmickienė I, Milinavičienė E, Everatt R. Hypertension, serum lipids and cancer risk: A review of epidemiological evidence. Medicina (Mex). 2016 //;52(2):89-98.
307. Sollars VE. Epigenetics as a Mechanism for Dietary Fatty Acids to Affect Hematopoietic Stem/Progenitor Cells And Leukemia—Royal Jelly for the Blood. Nutrition and Cancer From Epdemiology to Biology. 2012:65.

308. Dixon JB. Lymphatic Lipid Transport: Sewer or Subway? Trends in endocrinology and metabolism: TEM. 2010 06/11;21(8):480-7. PubMed PMID: PMC2914116.
309. Beguin Y, Aapro M, Ludwig H, Mizzen L, Osterborg A. Epidemiological and nonclinical studies investigating effects of iron in carcinogenesis--a critical review. Crit Rev Oncol Hematol. 2014 Jan;89(1):1-15. PubMed PMID: 24275533.
310. Ganz T, Nemeth E, editors. Iron sequestration and anemia of inflammation. Semin Hematol; 2009: Elsevier.
311. Richie JP, Kleinman W, Marina P, Abraham P, Wynder EL, Muscat JE. Blood iron, glutathione, and micronutrient levels and the risk of oral cancer. Nutr Cancer. 2008;60(4):474-82.
312. Stoltzfus RJ. Iron deficiency: global prevalence and consequences. Food and nutrition bulletin. 2003;24(4_suppl2):S99-S103.
313. Shibuya K, Mathers CD, Boschi-Pinto C, Lopez AD, Murray CJ. Global and regional estimates of cancer mortality and incidence by site: II. Results for the global burden of disease 2000. BMC Cancer. 2002;2(1):37.
314. Wang M, Spiegelman D, Kuchiba A, Lochhead P, Kim S, Chan AT, et al. Statistical methods for studying disease subtype heterogeneity. Stat Med. 2016 Feb 28;35(5):782-800. PubMed PMID: 26619806. Pubmed Central PMCID: 4728021. Epub 2015/12/02. eng.
315. Hollowell JG, Staehling NW, Flanders WD, Hannon WH, Gunter EW, Spencer CA, et al. Serum TSH, T4, and thyroid antibodies in the United States population (1988 to 1994): National Health and Nutrition Examination Survey (NHANES III). The Journal of Clinical Endocrinology & Metabolism. 2002;87(2):489-99.
316. Singer EA, Palapattu GS, van Wijngaarden E. Prostate-specific antigen levels in relation to consumption of nonsteroidal anti-inflammatory drugs and acetaminophen: results from the 2001-2002 National Health and Nutrition Examination Survey. Cancer. 2008 Oct 15;113(8):2053-7. PubMed PMID: 18780337. English.
317. Toth PP, Potter D, Ming EE. Prevalence of lipid abnormalities in the United States: the National Health and Nutrition Examination Survey 2003-2006. Journal of clinical lipidology. 2012 Jul-Aug;6(4):325-30. PubMed PMID: 22836069. Epub 2012/07/28. eng.
318. Menke A, Muntner P, Fernandez-Real JM, Guallar E. The association of biomarkers of iron status with mortality in US adults. Nutrition, metabolism, and cardiovascular diseases : NMCD. 2012 Sep;22(9):734-40. PubMed PMID: 21330119. Pubmed Central PMCID: PMC3138865. Epub 2011/02/19. eng.
319. Topic A, Djukic M. Diagnostic characteristics and application of alcohol biomarkers. Clin Lab. 2013;59(3-4):233-45. PubMed PMID: 23724610. Epub 2013/06/04. eng.
320. Kessenich CR, Cronin K. GGT and alcohol consumption. Nurse Pract. 2012 Aug 10;37(8):9-11. PubMed PMID: 22842136. Epub 2012/07/31. eng.
321. Grootendorst MR, Fitzgerald AJ, de Koning SGB, Santaolalla A, Portieri A, Van Hemelrijck M, et al. Use of a handheld terahertz pulsed imaging device to differentiate benign and malignant breast tissue. Biomedical Optics Express. 2017 Jun 1;8(6):2932-45. PubMed PMID: WOS:000404737200012. English.

322. Pereira J, Jeevan R, Browne J. First Annual Report of the National Mastectomy and Breast Reconstruction Audit. Department of Health; 2008.
323. Sakorafas GH. Breast cancer surgery--historical evolution, current status and future perspectives. Acta Oncol. 2001;40(1):5-18. PubMed PMID: 11321660.
324. Fitzal F, Gnant M. Breast Conservation: Evolution of Surgical Strategies. The Breast Journal. 2006 Sep-Oct;12(5 Suppl 2):S165-S73. PubMed PMID: 16958997.
325. Franceschini G, Magno S, Fabbri C, Chiesa F, Di Leone A, Moschella F, et al. Conservative and radical oncoplastic approches in the surgical treatment of breast cancer. Eur Rev Med Pharmacol Sci. 2008 Nov-Dec;12(6):387-96. PubMed PMID: 19146201. Epub 2009/01/17. eng.
326. Jeevan R, Cromwell D, Trivella M, Lawrence G, Kearins O, Pereira J, et al. Reoperation rates after breast conserving surgery for breast cancer among women in England: retrospective study of hospital episode statistics. BMJ. 2012;345:e4505.
327. Landercasper J, Whitacre E, Degnim AC, Al-Hamadani M. Reasons for re-excision after lumpectomy for breast cancer: insight from the american society of breast surgeons masterysm database. Ann Surg Oncol. 2014;21(10):3185-91.
328. Xue D, Qian C, Yang L, Wang X. Risk factors for surgical site infections after breast surgery: a systematic review and meta-analysis. European Journal of Surgical Oncology (EJSO). 2012;38(5):375-81.
329. Arora D, Hasan S, Male E, Abid R, Ord C, Dauway E. Cost analysis of re-excisions for breast conserving surgery in Central Texas. American Society of Clinical Oncology; 2015.
330. Sukumaran R, Somanathan T, Mathews A, Kattor J, Sambasivan S, Nair RP. Role of Frozen Section in Intraoperative Assessment of Ovarian Masses: a Tertiary Oncology Center Experience. Indian Journal of Surgical Oncology. 2014 04/11
01/27/received
03/24/accepted;5(2):99-103. PubMed PMID: PMC4116551.
331. Esbona K, Li Z, Wilke L. Intraoperative Imprint Cytology and Frozen Section Pathology for Margin Assessment in Breast Conservation Surgery: A Systematic Review. Ann Surg Oncol. 2012 07/31;19(10):3236-45. PubMed PMID: PMC4247998.
332. De Silva I, Rozen WM, Ramakrishnan A, Mirkazemi M, Baillieu C, Ptasznik R, et al. Achieving Adequate Margins in Ameloblastoma Resection: The Role for Intra-Operative Specimen Imaging. Clinical Report and Systematic Review. PLoS ONE. 2012 10/19
04/18/received
09/18/accepted;7(10):e47897. PubMed PMID: PMC3477138.
333. Sastry R, Bi WL, Pieper S, Frisken S, Kapur T, Wells W, et al. Applications of Ultrasound in the Resection of Brain Tumors. Journal of neuroimaging : official journal of the American Society of Neuroimaging. 2017 08/19;27(1):5-15. PubMed PMID: PMC5226862.
334. Birtoiu IA, Rizea C, Togoe D, Munteanu RM, Micsa C, Rusu MI, et al. Diagnosing clean margins through Raman spectroscopy in human and animal mammary tumour surgery: a short review. Interface Focus. 2016;6(6):20160067. PubMed PMID: PMC5071821.
335. Boerckel JD, Mason DE, McDermott AM, Alsberg E. Microcomputed tomography: approaches and applications in bioengineering. Stem Cell Research & Therapy. 2014 12/29;5(6):144. PubMed PMID: PMC4290379.

336. Calligaris D, Norton I, Feldman DR, Ide JL, Dunn IF, Eberlin LS, et al. Mass Spectrometry Imaging as a Tool for Surgical Decision-Making. Journal of mass spectrometry : JMS. 2013;48(11):1178-87. PubMed PMID: PMC3957233.
337. St John ER, Al-Khudairi R, Ashrafian H, Athanasiou T, Takats Z, Hadjiminas DJ, et al. Diagnostic Accuracy of Intraoperative Techniques for Margin Assessment in Breast Cancer Surgery: A Meta-analysis. Ann Surg. 2017 Feb;265(2):300-10. PubMed PMID: 27429028.
338. Bu L, Shen B, Cheng Z. Fluorescent imaging of cancerous tissues for targeted surgery. Adv Drug Del Rev. 2014 07/24;0:21-38. PubMed PMID: PMC4169718.
339. Arnone DD, Ciesla CM, Corchia A, Egusa S, Pepper M, Chamberlain JM, et al., editors. Applications of terahertz (THz) technology to medical imaging. Proc SPIE; 1999.
340. Smye SW, Chamberlain JM, Fitzgerald AJ, Berry E. The interaction between Terahertz radiation and biological tissue. Phys Med Biol. 2001 Sep;46(9):R101-12. PubMed PMID: 11580188. Epub 2001/10/03. eng.
341. Fitzgerald AJ, Wallace VP, Jimenez-Linan M, Bobrow L, Pye RJ, Purushotham AD, et al. Terahertz pulsed imaging of human breast tumors. Radiology. 2006 May;239(2):533-40. PubMed PMID: 16543586. Epub 2006/03/18. eng.
342. Fan S, He Y, Ung BS, Pickwell-MacPherson E. The growth of biomedical terahertz research. J Phys D: Appl Phys. 2014;47(37):374009.
343. Fitzgerald AJ, Wallace VP, Pye R, Jimenez-Linan M, Bobrow L, Purushotham AD, et al., editors. Terahertz imaging of breast cancer, a feasibility study. Infrared and Millimeter Waves, 2004 and 12th International Conference on Terahertz Electronics, 2004 Conference Digest of the 2004 Joint 29th International Conference on; 2004 27 Sept.-1 Oct. 2004.
344. Ashworth PC, Pickwell-MacPherson E, Provenzano E, Pinder SE, Purushotham AD, Pepper M, et al. Terahertz pulsed spectroscopy of freshly excised human breast cancer. Opt Express. 2009 Jul 20;17(15):12444-54. PubMed PMID: 19654646. Epub 2009/08/06. eng.
345. Jones DP. Sequencing the exposome: A call to action. Toxicology Reports. 2016 //;3:29-45.
346. Sheikh M, Coolen A. Accurate Bayesian Data Classification without Hyperparameter Cross-validation. arXiv preprint arXiv:171209813. 2017.

Appendix I

The use of Latent Class Analysis to predict cancer risk and cancer survival based on serum biomarkers

A Santaolalla[1], A Grigoriadis[2], AC Coolen[3], L Holmberg[1], H Garmo[1], N Hammar[4,5], H Malmstrom[4], G Walldius[6], I Jungner[7], M Van Hemelrijck[1]

1. King's College London, Division of Cancer Studies, Cancer Epidemiology Group, London; 2. King's College London, Division of Cancer Studies, Breast Cancer Now Research Unit, London; 3. King's College London, Institute for Mathematical and Molecular Biomedicine, London; 4. Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; 5. AstraZeneca Sverige, R&D, Gothenborg, Sweden; 6. Unit of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; 7. Department of Medicine, Clinical Epidemiological Unit, Karolinska Institutet and CALAB Research, Stockholm, Sweden

Introduction

- A wide variety of biomarkers have been studied, mainly independently, in relation to the risk of cancer
- There is still a need to understand the heterogeneity of biomarker patterns within the population to ultimately establish a panel of markers that could predict population at high risk of cancer
- We aim to identify biomarkers linked to carcinogenesis and their intrinsic associations by characterising specific subgroups of individuals based on serum biomarker profiles and their association with cancer risk and cancer survival

Study population

- The Swedish Apolipoprotein Mortality Risk Study (AMORIS)
- A total of 13,615 of individuals aged >20
- Baseline measurements of serum markers that are representative of the main metabolic pathways, determined by standard blood tests taken 3 years before any cancer diagnosis if occurred

Figure 1. Study data model. 16 biomarkers representing the main metabolic pathways were included in the LCA analysis. Outcome information, comorbidities and demographics were included in the Cox regression analysis together with classes obtained in the LCA

LCA Classes: Lipids, Liver and Iron markers defined the classes in our data model

- Our data analysis revealed four latent classes based on the model fit indicator Chi-square. Chi–square reached a minimum at four classes when running the null model on our data (Figure 3)
- LCA subgrouped our study population into 4 different biomarker profiles based on clinical cut-offs: one showing abnormal values for lipid markers (dyslipidemia), one presenting abnormal values for liver markers, one characterised for abnormal values of iron markers and one resembling the healthy population (Table 1)

Figure 3. Line-graph depicting the goodness of fit indicators (X^2(1) (Chi-square). The model fits best with four latent classes as determined by the minimum value reached by Chi-square

Classes	class 1	class 2	class 3	class 4
%	60	23	9	8
Biological interpretation	Normal	Lipids	Liver	Iron

Table 1. Estimated class population shares for the four different LCA classes obtained for our data model and the biological interpretation of the 4 different biomarker profiles

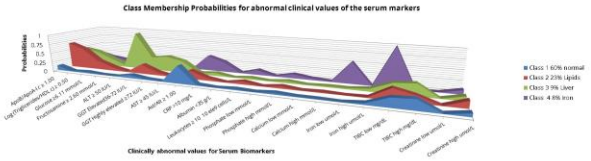


Figure 4. Class Membership Probabilities for abnormal clinical values of the serum markers for the four LCA classes. The four different biomarker profiles are represented in the graph

LCA Classes as cancer predictors

- A total of 3,158 (23.20%) individuals died, 14.37 % had a cancer diagnosis and 10.99% had circulatory events during a mean follow-up of 16.6 years
- Among all the deaths, 723 (5.31%) had cancer as the primary cause of death
- Individuals of the liver class presented worse overall survival, cancer-specific survival and risk of cancer
- Lipid marker profile showed positive association for both overall and cancer specific survival
- Associations were stronger when assessing overall survival for both classes
- Individuals belonging to the iron class do not present associations with any of the outcomes assessed

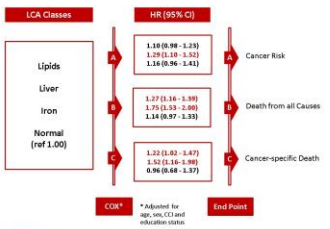


Figure 5. Hazard ratios and 95% confidence intervals for cancer risk (A), death from all causes (B) and cancer-specific death (C) for LCA classes in comparison to normal class as the reference

Statistical Analysis

- **Latent Class Analysis (LCA),** a data reduction method, was applied using our data model from AMORIS to characterise different classes of individuals based on their biomarker profiles, to ultimately evaluate intrinsic associations between those biomarkers. To run the LCA, all serum markers were dichotomised based on their clinical cut-offs
- **Multivariable Cox regression** was applied to examine whether the LCA classes based on the panel of serum biomarkers predicts risk of cancer and risk of all-cause and cancer-specific death

Figure 2. Statistical pipeline. LCA is applied on the data model obtaining a number of classes that are used as input models for the multivariable Cox regression analysis with different endpoints

Conclusions

- **Our findings present a new methodological approach to assess the association between multiple biomarkers and cancer diagnosis and survival in a well-defined population**
- **Latent Class Analysis showed that lipid metabolism plays an important role in classifying individuals based on their serum biomarkers, characterising 23% of the population; liver function and iron marker profiles distinguish important subgroups within the population**
- **Out of the four latent classes identified in our analysis, the markers of liver function manifested the strongest association with worse survival and cancer risk, suggesting a key role in the development of the disease and a potentially important area of research of mechanisms and clinical relevance**
- The lipid metabolism also showed correlation with overall survival and cancer survival
- We are currently working on validating these results in a different study population, using the NHANES cohort

KING'S College LONDON
Karolinska Institutet
Cancerfonden
Gunnar and Ingmar Jungner Foundation for Laboratory Medicine
NCRI National Cancer Research Institute

Appendix II

Big Data in Biology and Health

Metabolic profiles as risk factors to predict long-term cancer and mortality: the use of latent class analysis

A Santaolalla[1], H Garmo[1], A Grigoriadis[2], N Hammar[3,4], I Jungner[5], G Walldius[6], L Holmberg[1], M Van Hemelrijck[1]

1. King's College London, School of Cancer and Pharmaceutical Sciences, Translational Oncology & Urology Research (TOUR), London; 2. King's College London, School of Cancer and Pharmaceutical Sciences, Breast Cancer Now Research Unit, London; 3. Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; 4. AstraZeneca Sverige, R&D, Gothenburg, Sweden; 5. Department of Medicine, Clinical Epidemiological Unit, Karolinska Institutet and CALAB Research, Stockholm, Sweden; 6. Unit of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden;

Background of study

- A wide variety of biomarkers have been studied, mainly independently, in relation to the risk of cancer and mortality.
- There is still a need to understand the heterogeneity of biomarker patterns within the population to ultimately establish a panel of markers that could explain susceptibility to disease and predict population at long-term risk of cancer and mortality.
- To explore disease susceptibility and improve population risk stratification, we aimed to identify metabolic profiles linked to carcinogenesis and mortality and their intrinsic associations by characterising specific subgroups of individuals based on serum biomarker measurements.

Study population

The Swedish Apolipoprotein Mortality Risk Study (AMORIS)
A total of 13,615 of individuals aged >20
Baseline measurements of serum markers that are representative of the main metabolic pathways, determined by standard blood tests taken 3 years before any cancer diagnosis if occurred

Figure 1 . Study data model. 16 biomarkers representing the main metabolic pathways were included in the LCA analysis. Outcome information, comorbidities and demographics were included in the Cox regression analysis together with classes obtained in the LCA

Methodology

Starting from the **Blood Exposome**, we selected a group of metabolites that then feed into the statistical pipeline:

- **Latent Class Analysis (LCA),** a data reduction method, was applied to characterise different classes of individuals based on their biomarker profiles, to ultimately evaluate intrinsic associations between those biomarkers. To run the LCA, all serum markers were dichotomised based on their clinical cut-offs
- **Multivariable Cox regression** was applied to examine whether the LCA classes based on the panel of serum biomarkers predicts long-term risk of cancer and risk of all-cause and cancer-specific death

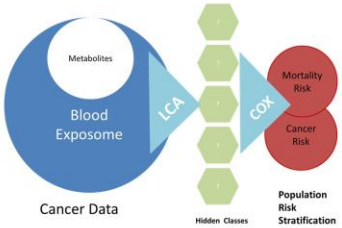


Figure 2 . Methodological approach. Innovative avenue to explore cancer susceptibility in a well-defined cohort. A shift from classical targeted hypothesis driven approach to a data driven.

LCA Classes: Lipids, Liver and Iron & Inflammation defined the metabolic profiles in our data model

Our data analysis revealed four latent classes based on the model fit indicator Chi-square. Chi–square reached a minimum at four latent classes when running models (null to 11 classes)(Figure 3)

LCA subgrouped our study population into 4 different biomarker profiles based on clinical cut-offs: one showing abnormal values for lipid markers (dyslipidemia), one presenting abnormal values for liver markers, one characterised for abnormal values of iron and inflammation markers and one resembling the healthy population (Figure 4)

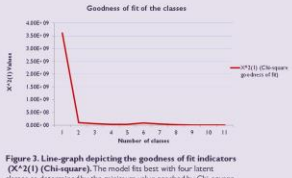


Figure 3. Line-graph depicting the goodness of fit indicators ($X^2(1)$ (Chi-square). The model fits best with four latent classes as determined by the minimum value reached by Chi-square.

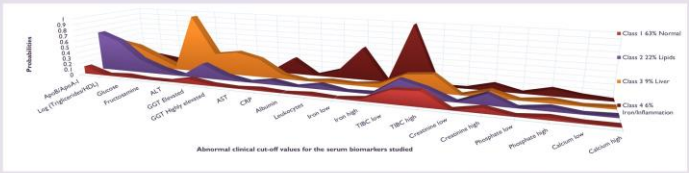


Figure 4. Class Membership Probabilities for abnormal clinical values of the serum markers for the four LCA –derived metabolic profiles .The four different biomarker profiles are represented in the graph.

LCA metabolic profiles as cancer predictors

A total of 3,158 (23.20%) individuals died, 14.37 % had a cancer diagnosis during a mean follow-up of 16.6 years

Individuals of the liver class presented worse cancer risk, overall death, cancer-specific and CDV death
Lipid marker profile showed positive association for both overall, cancer specific and CDV death
Iron/Inflammation profile showed association for both overall and CDV death

	HR (95% CI) time scale adjusted for age, sex and CCI
Cancer Risk : All cancer types	
Number of events	1956
1 - Normal class	1.00 (ref)
2 - Lipids	1.09 (0.98 - 1.22)
3 - Liver	1.78 (1.35 - 1.59)
4 - Iron/Inflammation	1.17 (0.97 - 1.41)

	HR (95% CI) time scale adjusted for age, sex and CCI
All causes death	
Number of events	3158
1 - Normal class	1.00 (ref)
2 - Lipids	1.26 (1.16 - 1.37)
3 - Liver	1.67 (1.47 - 1.90)
4 - Iron/Inflammation	1.21 (1.05 - 1.41)
Cancer death	
Number of events	706
1 - Normal class	1.00 (ref)
2 - Lipids	1.22 (1.02 - 1.45)
3 - Liver	1.44 (1.11 - 1.86)
4 - Iron/Inflammation	0.93 (0.66 - 1.32)
Cardiovascular death	
Number of events	1339
1 - Normal class	1.00 (ref)
2 - Lipids	1.39 (1.22 - 1.57)
3 - Liver	1.55 (1.25 - 1.93)
4 - Iron/Inflammation	1.29 (1.04 - 1.61)

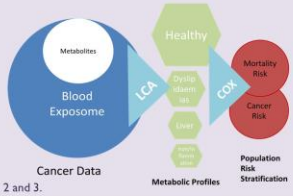
Figure 5. Hazard ratios and 95% confidence intervals for cancer risk, death from all causes, cancer-specific and cardiovascular death for LCA classes in comparison to normal class as the reference

Summary

We explored the blood exposome using metabolic markers of the population to assess how population heterogeneity is associated with cancer risk and mortality.

Using a data reduction method LCA that has characterised the population in 4 subgroups defined by metabolic profiles, healthy population, one abnormal lipids values population, one abnormal liver function and abnormal iron/inflammation.

These LCA derived metabolic profiles have shown using Cox regression analysis that all increase cancer risk and overall and cardiovascular mortality using the healthy class as a reference and worse cancer mortality for class 2 and 3.



Conclusions

The LCA adapted in this study illustrates how a biomarker-wide approach can help stratify a well defined population for cancer risk and mortality using a non-invasive standard diagnostic test

Systemic approaches assessing multiple types of biomarkers, as part of the exposome, at different time points in one's individual life, will help us to understand the complexity of cancer and better stratify population at risk of cancer and mortality

Gunnar and Ingmar Jungner Foundation for Laboratory Medicine
toûr
Cancerfonden
King's College London
Karolinska Institutet

Appendix III

operative radiotherapy. Consideration should be given to induction ILP in younger patients, who are likely to benefit most from avoiding radiotherapy.

http://dx.doi.org/10.1016/j.ejso.2015.08.033

3. Pulmonary primary and metastatic sarcomas – prognosis and post-resection survival – a 17 year analysis

Mohammad Salmasi, Mohammed Chowdhury, Robert Ashford, Catherine Richards, Apostolos Nakas, David Waller, Sridhar Rathinam
Glenfield Hospital, Leicester, UK

Introduction: Soft tissue sarcomas are rare neoplasms. Nearly one fifth of patients with primary sarcomas elsewhere will develop metastases during the course of their disease, most commonly to the lungs. Long term survival is possible in selected patients, particularly when pulmonary disease is resected.

Methods: We retrospectively analysed data at a single tertiary thoracic centre serviced by a regional sarcoma multidisciplinary meeting over a 17 year period (1996–2013). All surgical patients with primary and metastatic lung sarcomas were included. N = 56, 32 males; 24 females. Mean age 50 (range 11–84).

Results: Metastases were mainly from long bones, uterus, striated muscle, liver, other bones and chest wall. Histological diagnoses were leiomyosarcomas 31%, osteosarcomas 14%, spindle cell tumour 10%, other 9%, unspecified 21%.

Operatively, 31% of patients had a wedge resection of tumour, 28% had an anatomical resection (ratio of lobectomy: pneumonectomy = 3:1) and 41% had a combination of wedge and anatomical resection. 46% of procedures were performed via video assisted thoracoscopic surgery (VATS), the remainder were open procedures. 34% of patients went on to require a second surgical resection.

4 patients (7%) died with 30 days post operatively (mean age 75.5). Overall median survival 20 months (SE 8.6, 95% CI 3.2–37.0). Kaplan-Meier analysis found 64% survival at 1-year, 49% at 2 years, 40% at 3 years, 35% at 4 years and 28% at 5 years.

Conclusion: Metastatectomy can be performed safely in our cohort with good outcomes. Age is a determinant of 30 day mortality. VATS metastectomy offers the ability of repeated metastectomy with minimal morbidity. Multi-modality treatment in this setting of a regional sarcoma MDT is valuable in this specialist area of thoracic surgery.

http://dx.doi.org/10.1016/j.ejso.2015.08.034

4. Morbidity, mortality and oncological outcome following multi-visceral resection of retroperitoneal sarcomas

Henry Smith, Deepa Panchalingam, Jonathan Hannay, Myles Smith, Joseph Meirion Thomas, Andrew Hayes, Dirk Strauss
The Royal Marsden Hospital NHS Foundation Trust, London, UK

Introduction: Retroperitoneal sarcoma (RPS) is not a single disease entity, but comprises a range of different histological subtypes with dissimilar behaviour and biology. This study sought to characterise the morbidity and mortality associated with multi-visceral resection and oncological outcomes according to subtype.

Methods: All patients undergoing resection of primary RPS at a single institution between January 2005 and January 2015 were identified from a kidney (190), colon (186), spleen (64) and pancreas (46). Thirty-day mortality was 1.4% (5/362). Thirty-day morbidity was 16.0% (58/362), with 25 patients requiring a return to theatre. Age over 75 years was a predictive factor for 30-day mortality, with tumour grade a predictor of 30-day morbidity. The overall anastomotic leak rate was 2.2% (4/186). Nephrectomy was performed in 190 patients, 112 had follow-up renal function >30 days post-operatively. Median pre-operative glomerular filtration rate was 89.6 ml/min falling to a nadir of 47.1 ml/min before recovering to 58.1 ml/min. Disease specific survival at 5-years for the whole cohort was 70.4%. For well-differentiated liposarcoma, dedifferentiated liposarcoma and leiomyosarcoma. 5-year local recurrence free survival was 77.4%, 35.1% and 80.3%, respectively. Distant metastasis-free survival was 100%, 83.5% and 61.7%, respectively. Disease specific survival was 97.2%, 65.8% and 63.9%, respectively.

Conclusion: Multi-visceral resection of RPS is associated with low morbidity when performed in a high-volume centre. Histological subtype and grade determines patterns of local and distant recurrence.

http://dx.doi.org/10.1016/j.ejso.2015.08.035

5. The use of a handheld Terahertz pulsed imaging device to differentiate benign and malignant breast tissue with a view to reducing re-operation rates in breast-conserving surgery

Maarten Grootendorst[1,2], Susan Brouwer de Koning[1,2], Tony Fitzgerald[3], Alessia Portieri[4], Aida Santaolalla[1], Massi Cariati[1,2], Michael Pepper[4,5], Vincent Wallace[3], Sarah Pinder[1], Arnie Purushotham[1,2]
[1] Division of Cancer Studies, King's College London, London, UK
[2] Guy's and St Thomas' NHS Foundation Trust, London, UK
[3] School of Physics, University of Western Australia, Perth, Australia
[4] Teraview Ltd., Cambridge, UK
[5] London Centre for Nanotechnology, University College London, London, UK

Introduction: This study evaluates the ability of Terahertz Pulsed Imaging (TPI) to discriminate benign from malignant breast tissue, with the aim of developing a technique for intraoperative tumour margin assessment to reduce the re-operation rate in breast-conserving surgery (BCS).

Methods: 46 breast tissue samples (30 patients) from freshly excised breast cancer specimens were scanned using a 0–2.0 THz handheld TPI probe (Teraview) (REC 12-EE-0493). For each sample, detailed pathology at 0.6mm-intervals was obtained and correlated with THz data. The performance of parameters from the full THz time-domain pulse to discriminate fibrous from tumour was quantitatively evaluated with an area under the receiver operating characteristic curve (AUROC) analysis. A Mann-Whitney test was performed on parameters with an AUROC value >0.75 to determine whether the parameter values were statistically significantly different.

Results: 16 invasive ductal carcinoma, 3 invasive lobular carcinoma, 1 invasive tubular carcinoma and 26 benign (adipose/fibrous) samples were used. Adipose tissue could be discriminated from tumour/fibrous tissue using the full time-domain pulse. Tumour could be discriminated from fibrous tissue using 5 parameters; amplitudes and integrals of time-domain pulse (AUROC $\geq$ 0.77; $p < 0.001$).

Conclusion: TPI can accurately discriminate benign from malignant tissue in an *ex vivo* setting, warranting *in vivo* evaluation to test the impact on re-excision rates in BCS.

http://dx.doi.org/10.1016/j.ejso.2015.08.036

Appendix IV

Research Article | Vol. 8, No. 6 | 1 Jun 2017 | BIOMEDICAL OPTICS EXPRESS 2932
Biomedical Optics EXPRESS

Use of a handheld terahertz pulsed imaging device to differentiate benign and malignant breast tissue

MAARTEN R. GROOTENDORST,[1,2,7] ANTHONY J. FITZGERALD,[3,7] SUSAN G. BROUWER DE KONING,[1,2] AIDA SANTAOLALLA,[1] ALESSIA PORTIERI,[4] MIEKE VAN HEMELRIJCK,[1] MATTHEW R. YOUNG,[2] JULIE OWEN,[5] MASSI CARIATI,[1,2] MICHAEL PEPPER,[4,6] VINCENT P. WALLACE,[3] SARAH E. PINDER,[5] AND ARNIE PURUSHOTHAM[1,2]

[1]*King's College London, Division of Cancer Studies, London, UK*
[2]*Department of Breast Surgery, Guy's and St Thomas' NHS Foundation Trust, London, UK*
[3]*School of Physics, University of Western Australia, Perth, Australia*
[4]*Teraview Ltd., Cambridge, UK*
[5]*King's College London, Division of Cancer Studies, King's Health Partners Cancer Biobank and Breast Pathology Research Group, London, UK;*
[6]*London Centre for Nanotechnology, University College London, UK*
[7]*Contributed equally*
**arniepurushotham@gmail.com*

Abstract: Since nearly 20% of breast-conserving surgeries (BCS) require re-operation, there is a clear need for developing new techniques to more accurately assess tumor resection margins intraoperatively. This study evaluates the diagnostic accuracy of a handheld terahertz pulsed imaging (TPI) system to discriminate benign from malignant breast tissue ex vivo. Forty six freshly excised breast cancer samples were scanned with a TPI handheld probe system, and histology was obtained for comparison. The image pixels on TPI were classified using (1) parameters in combination with support vector machine (SVM) and (2) Gaussian wavelet deconvolution in combination with Bayesian classification. The results were an accuracy, sensitivity, specificity of 75%, 86%, 66% for method 1, and 69%, 87%, 54% for method 2 respectively. This demonstrates the probe can discriminate invasive breast cancer from benign breast tissue with an encouraging degree of accuracy, warranting further study.

OCIS codes: (110.6795) Terahertz imaging; (170.3880) Medical and biological imaging; (170.6935) Tissue characterization.

References and links

1. J. F. I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, and D. Forman, "GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11" (International Agency for Research on Cancer. World Health Organisation, 2012).
2. R. Jeevan, J. Browne, J. Van der Meulen, J. Pereira, C. Caddy, C. Sheppard, C. McGregor-Johnson, Z. Kramer, and S. Dead, "First Annual Report of the National Mastectomy and Breast Reconstruction Audit 2008" (2008).
3. K. P. McGuire, A. A. Santillan, P. Kaur, T. Meade, J. Parbhoo, M. Mathias, C. Shamehdi, M. Davis, D. Ramos, and C. E. Cox, "Are mastectomies on the rise? A 13-year trend analysis of the selection of mastectomy versus breast conservation therapy in 5865 patients," Ann. Surg. Oncol. **16**(10), 2682–2690 (2009).
4. R. Jeevan, D. A. Cromwell, M. Trivella, G. Lawrence, O. Kearins, J. Pereira, C. Sheppard, C. M. Caddy, and J. H. van der Meulen, "Reoperation rates after breast conserving surgery for breast cancer among women in England: retrospective study of hospital episode statistics," BMJ **345**, e4505 (2012).
5. J. Landercasper, E. Whitacre, A. C. Degnim, and M. Al-Hamadani, "Reasons for re-excision after lumpectomy for breast cancer: insight from the American Society of Breast Surgeons Mastery(SM) database," Ann. Surg. Oncol. **21**(10), 3185–3191 (2014).
6. D. Q. Xue, C. Qian, L. Yang, and X. F. Wang, "Risk factors for surgical site infections after breast surgery: a systematic review and meta-analysis," Eur. J. Surg. Oncol. **38**, 375–381 (2012).

#286058 https://doi.org/10.1364/BOE.8.002932
Journal © 2017 Received 9 Feb 2017; accepted 2 May 2017; published 9 May 2017

Biomedical Optics EXPRESS

7. J. Heil, K. Breitkreuz, M. Golatta, E. Czink, J. Dahlkamp, J. Rom, F. Schuetz, M. Blumenstein, G. Rauch, and C. Sohn, "Do reexcisions impair aesthetic outcome in breast conservation surgery? Exploratory analysis of a prospective cohort study," Ann. Surg. Oncol. **19**(2), 541–547 (2012).
8. D. Arora, S. Hasan, E. Male, R. Abid, C. Ord, E. Dauway, and B. Scott, "Cost analysis of re-excisions for breast conserving surgery in Central Texas," in *ASCO Annual Meeting* (2015)
9. E. R. St John, R. Al-Khudairi, H. Ashrafian, T. Athanasiou, Z. Takats, D. J. Hadjiminas, A. Darzi, and D. R. Leff, "Diagnostic Accuracy of Intraoperative Techniques for Margin Assessment in Breast Cancer Surgery: A Meta-analysis," Ann. Surg. **265**(2), 300–310 (2016).
10. K. Deng, C. Zhu, X. Ma, H. Jia, Z. Wei, Y. Xiao, and J. Xu, "Rapid Discrimination of Malignant Breast Lesions from Normal Tissues Utilizing Raman Spectroscopy System: A Systematic Review and Meta-Analysis of In Vitro Studies," PLoS One **11**(7), e0159860 (2016).
11. J. Q. Brown, T. M. Bydlon, S. A. Kennedy, M. L. Caldwell, J. E. Gallagher, M. Junker, L. G. Wilke, W. T. Barry, J. Geradts, and N. Ramanujam, "Optical spectral surveillance of breast tissue landscapes for detection of residual disease in breast tumor margins," PLoS One **8**(7), e69906 (2013).
12. L. G. Wilke, J. Q. Brown, T. M. Bydlon, S. A. Kennedy, L. M. Richards, M. K. Junker, J. Gallagher, W. T. Barry, J. Geradts, and N. Ramanujam, "Rapid noninvasive optical imaging of tissue composition in breast tumor margins," Am. J. Surg. **198**(4), 566–574 (2009).
13. M. D. Keller, S. K. Majumder, M. C. Kelley, I. M. Meszoely, F. I. Boulos, G. M. Olivares, and A. Mahadevan-Jansen, "Autofluorescence and diffuse reflectance spectroscopy and spectral imaging for breast surgical margin analysis," Lasers Surg. Med. **42**(1), 15–23 (2010).
14. A. M. Laughney, V. Krishnaswamy, E. J. Rizzo, M. C. Schwab, R. J. Barth, Jr., D. J. Cuccia, B. J. Tromberg, K. D. Paulsen, B. W. Pogue, and W. A. Wells, "Spectral discrimination of breast pathologies in situ using spatial frequency domain imaging," Breast Cancer Res. **15**(4), R61 (2013).
15. I. J. Bigio, S. G. Bown, G. Briggs, C. Kelley, S. Lakhani, D. Pickard, P. M. Ripley, I. G. Rose, and C. Saunders, "Diagnosis of breast cancer using elastic-scattering spectroscopy: preliminary clinical results," J. Biomed. Opt. **5**(2), 221–228 (2000).
16. A. M. Zysk, K. Chen, E. Gabrielson, L. Tafra, E. A. May Gonzalez, J. K. Canner, E. B. Schneider, A. J. Cittadine, P. Scott Carney, S. A. Boppart, K. Tsuchiya, K. Sawyer, and L. K. Jacobs, "Intraoperative Assessment of Final Margins with a Handheld Optical Imaging Probe During Breast-Conserving Surgery May Reduce the Reoperation Rate: Results of a Multicenter Study," Ann. Surg. Oncol. **22**(10), 3356–3362 (2015).
17. F. T. Nguyen, A. M. Zysk, E. J. Chaney, J. G. Kotynek, U. J. Oliphant, F. J. Bellafiore, K. M. Rowland, P. A. Johnson, and S. A. Boppart, "Intraoperative Evaluation of Breast Tumor Margins with Optical Coherence Tomography," Cancer Res. **69**(22), 8790–8796 (2009).
18. S. J. Erickson-Bhatt, R. M. Nolan, N. D. Shemonski, S. G. Adie, J. Putney, D. Darga, D. T. McCormick, A. J. Cittadine, A. M. Zysk, M. Marjanovic, E. J. Chaney, G. L. Monroy, F. A. South, K. A. Cradock, Z. G. Liu, M. Sundaram, P. S. Ray, and S. A. Boppart, "Real-time imaging of the resection bed using a handheld probe to reduce incidence of microscopic positive margins in cancer surgery," Cancer Res. **75**(18), 3706–3712 (2015).
19. J. Balog, L. Sasi-Szabó, J. Kinross, M. R. Lewis, L. J. Muirhead, K. Veselkov, R. Mirnezami, B. Dezső, L. Damjanovich, A. Darzi, J. K. Nicholson, and Z. Takáts, "Intraoperative tissue identification using rapid evaporative ionization mass spectrometry," Sci. Transl. Med. **5**(194), 194ra93 (2013).
20. E. R. St John, R. Al-Khudairi, J. Balog, M. Rossi, L. Gildea, A. Speller, R. Ramakrishnan, S. Shousha, Z. Takats, D. R. Leff, and A. Darzi, "Rapid evaporative ionisation mass spectrometry towards real time intraoperative oncological margin status determination in breast conserving surgery.," in *38th Annual San Antonio Breast Cancer Symposium*, (San Antonio, 2016).
21. J. M. Dixon, L. Renshaw, O. Young, D. Kulkarni, T. Saleem, M. Sarfaty, R. Sreenivasan, C. Kusnick, J. Thomas, and L. J. Williams, "Intra-operative assessment of excised breast tumour margins using ClearEdge imaging device," Eur. J. Surg. Oncol. **42**(12), 1834–1840 (2016).
22. A. L. Vahrmeijer, M. Hutteman, J. R. van der Vorst, C. J. van de Velde, and J. V. Frangioni, "Image-guided cancer surgery using near-infrared fluorescence," Nat. Rev. Clin. Oncol. **10**(9), 507–518 (2013).
23. C. Yu, S. Fan, Y. Sun, and E. Pickwell-Macpherson, "The potential of terahertz imaging for cancer diagnosis: A review of investigations to date," Quant. Imaging Med. Surg. **2**(1), 33–45 (2012).
24. S. Fan, Y. He, B. S. Ung, and E. Pickwell-MacPherson, "The growth of biomedical terahertz research," J. Phys. D Appl. Phys. **47**(37), 374009 (2014).
25. A. J. Fitzgerald, V. P. Wallace, M. Jimenez-Linan, L. Bobrow, R. J. Pye, A. D. Purushotham, and D. D. Arnone, "Terahertz pulsed imaging of human breast tumors," Radiology **239**(2), 533–540 (2006).
26. P. C. Ashworth, E. Pickwell-MacPherson, E. Provenzano, S. E. Pinder, A. D. Purushotham, M. Pepper, and V. P. Wallace, "Terahertz pulsed spectroscopy of freshly excised human breast cancer," Opt. Express **17**(15), 12444–12454 (2009).
27. A. J. Fitzgerald, S. Pinder, A. D. Purushotham, P. O'Kelly, P. C. Ashworth, and V. P. Wallace, "Classification of terahertz-pulsed imaging data from excised breast tissue," J. Biomed. Opt. **17**(1), 016005 (2012).
28. C. C. Park, M. Mitsumori, A. Nixon, A. Recht, J. Connolly, R. Gelman, B. Silver, S. Hetelekidis, A. Abner, J. R. Harris, and S. J. Schnitt, "Outcome at 8 years after breast-conserving surgery and radiation therapy for invasive breast cancer: influence of margin status and systemic therapy on local recurrence," J. Clin. Oncol. **18**(8), 1668–1675 (2000).

29. K. S. Khan and P. F. Chien, "Evaluation of a clinical test. I: assessment of reliability," BJOG **108**(6), 562–567 (2001).
30. A. Shalabi, M. Inoue, J. Watkins, E. De Rinaldis, and A. C. Coolen, "Bayesian clinical classification from high-dimensional data: Signatures versus variability," Stat. Methods Med. Res. DOI: 10.1177/0962280216628901 (2016).
31. M. Ahmed, I. T. Rubio, J. M. Klaase, and M. Douek, "Surgical treatment of nonpalpable primary invasive and in situ breast cancer," Nat. Rev. Clin. Oncol. **12**(11), 645–663 (2015).
32. P. C. Ashworth, P. O'Kelly, A. D. Purushotham, S. E. Pinder, M. Kontos, M. Pepper, and V. P. Wallace, "An intra-operative THz probe for use during the surgical removal of breast tumors," in *Infrared, Millimeter and Terahertz Waves, 2008. IRMMW-THz 2008. 33rd International Conference on*, 2008), 1–3.
33. B. C. Truong, H. D. Tuan, A. J. Fitzgerald, V. P. Wallace, and H. T. Nguyen, "A dielectric model of human breast tissue in terahertz regime," IEEE Trans. Biomed. Eng. **62**(2), 699–707 (2015).
34. P. C. Stomper, D. J. D'Souza, P. A. DiNitto, and M. A. Arredondo, "Analysis of parenchymal density on mammograms in 1353 women 25-79 years old," AJR Am. J. Roentgenol. **167**(5), 1261–1265 (1996).

1. Introduction

Breast cancer is by far the most common cancer among women worldwide [1]. A combination of an increased use of screening mammography, neoadjuvant chemotherapy and neoadjuvant endocrine therapy to downstage the size of the tumor, has significantly increased the number of patients suitable for breast-conserving surgery (BCS). Currently approximately two-thirds of newly diagnosed breast cancer patients in the United Kingdom and the United States undergo BCS as initial treatment [2, 3].

A key problem in BCS is that on average approximately 20% of patients require a re-operation because of close or positive tumor margins on postoperative histopathological analysis [4, 5]. Re-operations potentially have a significant impact on patients and healthcare systems. They can result in an increased rate of surgical complications [6], compromise cosmetic outcome [7], delay adjuvant therapy, and increase anxiety and stress for patients and their families. Re-excision surgery also presents a high cost burden to healthcare systems; a recent study in the United States showed that the costs for a re-excision was $4721 per patient [8].

In an attempt to decrease the re-operation rate, techniques to intraoperatively assess tumor resection margins have been developed. Clinically established techniques include specimen radiography, intraoperative ultrasound, radiofrequency spectroscopy, frozen section analysis and touch imprint cytology. However, these all have limitations in terms of diagnostic accuracy, logistical or technical demands or cost-effectiveness [9]. Emerging techniques include Raman spectroscopy [10], diffuse reflectance spectroscopy [11–15], optical coherence tomography [16–18], mass spectroscopy [19, 20], bio-impedance spectroscopy [21], and (targeted) fluorescence imaging [22]. These techniques each have unique limitations, and their potential value for improving quality of care and reducing healthcare costs is yet unknown.

Terahertz pulsed imaging (TPI) employs terahertz (THz) radiation (0.1 – 4 THz) for imaging biological tissue. Due to the millimetric penetration depth and sensitivity of THz radiation to changes in water content and tissue composition, and the submillimeter imaging resolution of TPI, this technique holds promise for imaging cancer [23]. Work performed to date has shown the ability of TPI to discriminate malignant from benign tissue in skin, colon, oral, gastric, brain and breast cancer [24]. In 2006, Fitzgerald *et al.* were the first to demonstrate the potential of TPI for identifying breast cancer during BCS [25]. They measured 22 freshly excised breast tissue samples, and demonstrated a good correlation between tumor size and shape assessed by TPI compared with histopathology. To better understand the origin of the observed contrast on TPI, Ashworth *et al.* used THz spectroscopy and showed that the absorption coefficient and refractive index of tumor were different to that of normal breast tissue in the THz region of the spectrum [26]. Following from their initial work, Fitzgerald *et al.* imaged 51 breast samples and assessed the diagnostic accuracy of TPI by using a range of THz image parameters and classification techniques [27]. They demonstrated an accuracy, sensitivity and specificity of 92%, 90% and 92%, respectively.

However, the TPI device used in their study is not suitable for intraoperative assessment of intact breast specimens due to the requirement for physical tissue disruption to obtain samples that fit the 20 x 20 mm sample holder. Importantly, the tissue samples included in their data set had a 'homogeneous' tissue composition, i.e. contained more than 50% of a single tissue type. This is not an accurate representation of the tissue composition found at the resection border of patients with close or positive margins, as involved margins are often identified microscopically as a small number of tumor cells immersed in a 'background' of fibrous and/or adipose tissue [28]. Thus, the diagnostic accuracy of TPI for detecting tumor close or at the margin remains underdetermined.

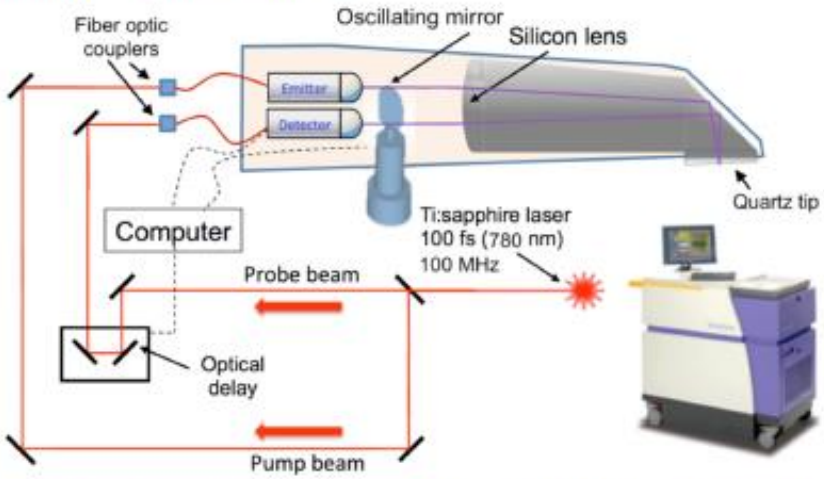


Fig. 1. Schematic illustration of TPI handheld probe system. The emitted laser pulses are split into a "pump beam" and a "probe beam". The pump beam is guided through the optical fibers in the umbilical cord, and subsequently incident on the photoconductive emitter to produce THz pulses. The probe beam is guided onto the photoconductive detector to detect the THz pulses reflected from the tissue sample. By altering the path length of the probe beam, the time of arrival at the detector in respect to the incident THz pulse can be changed, thus sampling the THz pulse in the time domain.

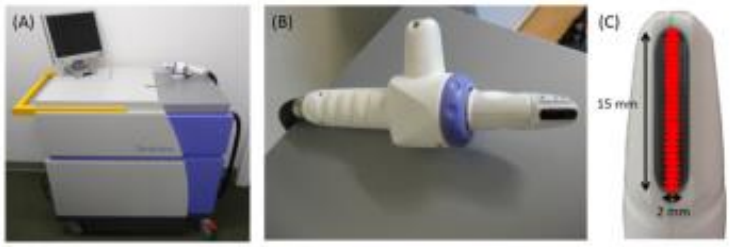


Fig. 2. TPI handheld probe system. (A) Main unit with computer monitor, handheld imaging probe and black umbilical cord (visible on the right). (B) Close up of the handheld imaging probe. (C) Close up of the head of the imaging probe showing the black quartz window. The probe scans an area of 15 x 2 mm, and acquires data from 26 pixels (red).

To facilitate the use of TPI to scan tumor resection margins intraoperatively, Teraview Ltd. (Cambridge, UK) has developed a *handheld* probe system. A single center study was performed to evaluate the ability of the TPI handheld probe to discriminate benign from malignant breast tissue in an *ex vivo* setting. The aims of the study were to obtain a data set that closely resembles the mixture of benign and tumor tissue commonly found at the resection border of patients with involved margins, and to evaluate the diagnostic

Biomedical Optics EXPRESS

performance of the TPI handheld probe in terms of accuracy, sensitivity, specificity and predictive values using two data analysis and classification methods.

2. Methods

2.1 TPI handheld probe system

The TPI handheld probe produces and detects THz pulses by guiding laser pulses from a femtosecond fiber laser (Menlo Systems GmbH, Martinsreid, Germany) down optical fibers to a photoconductive emitter and detector (Fig. 1). The 0.1 – 1.8 THz pulses are then guided by an oscillating mirror via a monolithic silicon section onto a quartz window present at the tip of the probe, scanning 26 pixels in an area of 15 x 2 mm at a frequency of 4 Hz (Fig. 2). During scanning each pixel acquires THz pulses over time to form a TPI image (Fig. 3).

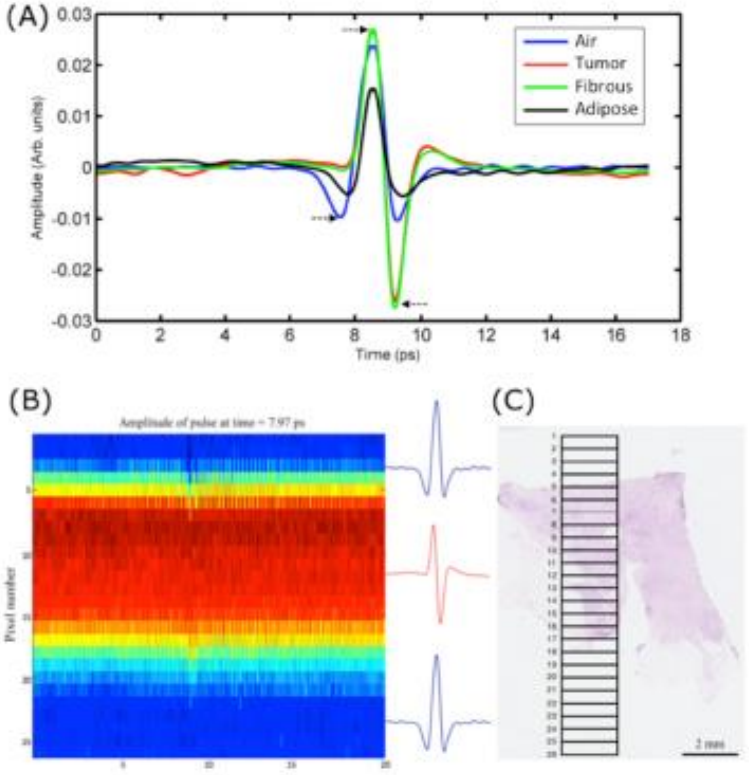


Fig. 3. Correlating TPI with histopathology. (A) Typical impulse function from breast tissue containing a high percentage of tumor, fibrous, and adipose cells, and air, respectively. Clear differences are seen between the impulse functions from air and from tissue, and between adipose and tumor/fibrous tissue, especially at time points t = 7.97 ps, t = 8.93 ps, and t = 9.67 ps (black arrows). (B) TPI image from sample based on the amplitude of the impulse function at t = 7.97ps. A clear contrast can be seen at the air-tissue interface at pixel 5 and pixel 17. Note an 'edge effect' at these interfaces, causing a distortion in the impulse functions of these pixels. (C) Digital histopathology slide of the same tissue sample. By using the photograph of the sample in combination with the air-tissue interface visible in the TPI image, the TPI 15 x 2 mm scan area can be accurately mapped onto the histopathology slide (black rectangle). The pixels are displayed as intermittent horizontal lines at 0.6 mm distance in the scan window. Pixel 5 – 17 contain invasive ductal/no special type (NST) carcinoma; the percentage of tumor cells in each of these pixel areas ranges between 5 – 10%. The tissue immediately surrounding the tumor cells (background) is composed of fibrous tissue, whilst fatty adipose tissue is seen inferiorly.

Biomedical Optics EXPRESS

2.2 Data acquisition

Between August 2013 and August 2014, following written, informed consent, breast tissue samples from patients who underwent BCS or mastectomy at Guy's Hospital in London were scanned with the TPI probe (REC 12-EE-0493). Within 60 minutes post-excision, BCS or mastectomy specimens were inked and sliced by an Advanced Practitioner in the King's Health Partners Cancer Biobank located adjacent to the operating theatre. Tissue samples were obtained for the study subject to the amount of tissue required for diagnostic purposes.

Prior to scanning the samples, a Tegaderm layer (3M Tegaderm Film, 3M, Bracknell, UK) was applied to the probe's quartz window, and the remainder of the probe was wrapped in a disposable protective sheath to prevent contamination from tissue. To enable consistent and controlled TPI measurements, tissue samples were placed in a standard histology cassette (Unisette, Simport, Beloeil, Canada) that tightly fitted the head of the probe (Fig. 4). All samples were scanned for 20 seconds. Upon completion of each measurement a photograph of the sample in the cassette was taken to facilitate accurate correlation of the TPI data with the final histology slide.

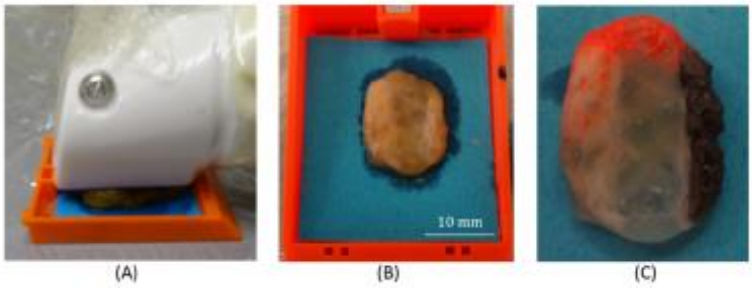


Fig. 4. TPI measurement of tissue sample. (A) TPI handheld probe measurement of tissue sample positioned in histology cassette. Note that the head of the imaging probe tightly fits in the cassette, which facilitates applying a consistent pressure throughout the measurement, while preventing displacement of the probe. (B) Photograph of the tissue sample obtained after the sample was scanned. The imprint of the scan window on the sample is clearly visible. This photograph was used to facilitate accurate correlation of TPI data with histopathology. (C) Photograph of tissue sample after it was inked. Inking was performed to enable spatial orientation of the sample when analysed microscopically by the pathologist.

After the sample was scanned, the Tegaderm layer was removed from the probe and a 60 second air measurement was performed that was used as a reference for data processing. For orientation purposes the top and the right surface of the sample were inked red and black respectively, after which the histology cassette containing the sample was closed, placed in formalin for 24-48 hours, and subsequently processed and paraffin wax embedded. Three to 4 micron sections were then cut and stained with hematoxylin and eosin. The histology slides were digitalized, and subsequently viewed and analyzed using histopathology slide viewer software (NDP.view2, Hamamatsu, UK).

2.3 TPI data processing

Each pixel of the TPI probe acquired raw THz pulses throughout the duration of the measurement. These pulses were deconvolved with the reference ("air without Tegaderm") pulses and a double Gaussian filter was applied to reduce noise. All pulses were aligned in time to compensate for small offsets in the phase of the detected pulses due to changes in the optical path length that occur when optical fibers deform slightly with movement during scanning. The deconvolved pulses – called impulse functions - of each pixel were then averaged over time, resulting in 26 impulse functions, one for each of the 26 pixels to be used for discriminating benign from malignant breast tissue (Fig. 3(B)).

Biomedical Optics EXPRESS

2.4 Correlation of TPI with histopathology

By using the photograph depicting the imprint of the probe's scan window on the sample, and the clear contrast from the air-tissue interface and tissue composition on the TPI image, the 15 x 2 mm TPI scan area was mapped onto the digital histopathology image (Fig. 3(C)). To reduce potential inaccuracies in correlating TPI with histopathology, samples were excluded from further analysis if the number of tissue-containing pixels on TPI and histopathology differed by more than three.

2.5 Histopathological analysis and selection of TPI data

The digital histopathology slide of each sample was analyzed in the viewer software by a Consultant breast histopathologist (S.E.P.), and the percentage of different tissue types, namely tumor, fibrous, and adipose, were recorded in 5% intervals. For a subset of samples an intra-rater variability analysis was performed to assess the ability of the histopathologist to consistently score the tissue samples. For this analysis, a total of 92 pixels from 7 tumor samples and 125 pixels from 7 benign samples were re-evaluated in a blinded method by the same histopathologist 8 weeks after the first analysis. Weighted kappa coefficients were calculated to assess the agreement in subgroup classification between evaluation 1 and 2 (*kappa 2* function of the 'irr' package v0.84, R statistical software v3.2.2). A kappa coefficient (κ) greater than 0.80 was considered excellent agreement [29]. A Wilcoxon signed-rank test was performed to assess whether evaluation 1 was statistically significantly different from evaluation 2. The level of significance was defined as $p < 0.05$.

Table 1. Pixel characteristics analysis data set. A total of 257 pixels were included in the TPI data set: 115 tumor pixels, 116 fibrous pixels and 26 pure adipose pixels. The tumor pixels predominantly consisted of invasive ductal/no special type carcinoma (N = 92) and invasive lobular carcinoma (N = 19). Most of the tumor pixels contained a low to moderate percentage of tumor cells ranging between 1 – 60% (N = 98). Almost all tumor cells had a background of pure fibrous tissue; only 5 had a background containing a mix of fibrous and adipose. Most of the fibrous pixels had a high percentage of fibrous cells ranging between 81 – 100% (N = 91). Only 26 of the 257 pixels consisted of pure adipose tissue.

	Tumor					Fibrous		Adipose
Tissue percentage groups (%)	NST	NST + DCIS	ILC	Number pixels	BG	Number pixels	BG	Number pixels
81 – 100	3	1		4	F	91	A	26*
61 – 80	11	2		13	F	2	A	
41 – 60	22		6	28	F	7	A	
21 – 40	33	1	12	46	F: 43 F/A: 3	3	A	
1 – 20	23		1	24	F: 22 F/A: 2	13	A	
Number pixels	92	4	19	115		116		26

NST = invasive ductal/no special type carcinoma; DCIS = ductal carcinoma *in situ*; ILC = invasive lobular carcinoma; BG = background tissue. In our dataset the background consisted of fibrous tissue (F), adipose tissue (A), or a mixture of fibrous and adipose tissue (F/A).

*These pixels contained 100% adipose tissue.

Biomedical Optics EXPRESS

2.6 TPI data analysis and classification

Classification of each of the selected impulse functions as malignant or benign was performed using two data analysis and classification methods: (1) heuristic parameters in combination with support vector machine (SVM) classification and (2) Gaussian wavelet deconvolution with Bayesian classification.

The impulse function of each pixel is made of values at 301 time points, and given that information from both the time and frequency domain can be used to classify pixels, it was advantageous to reduce the dimensionality of the data for classification. In method 1 this was done by using parameters that described significant features in the impulse function or spectrum. Since a large number of time points or frequency points can be selected to form a parameter, a Receiver operator Characteristics (ROC) analysis was used to select the optimal characterizing parameters. The area under the ROC curve (AUROC) was used as an estimate of the classification ability of each parameter (illustrated in Fig. 5). AUROC analysis was performed on the data from 3 pathology groups; (i) the entire tumor and fibrous groups (excluding pure adipose), (ii) the tumor and 100% fibrous groups, and (iii) the tumor and 100% adipose groups. From this analysis the top 11 parameters were selected to be used for classification of the data in the SVM (Table 2). To avoid effects of overfitting, the parameters chosen were tested to ensure they were not correlated, eliminating any parameters with an absolute correlation coefficient of 0.7 or more. The SVM function used for classification was from the Matlab native functions *svmtrain.m* and *svmclassify.m* using a radial basis function as the kernel to form the decision boundaries (MATLAB 2013A, The Mathworks Inc., Natick, MA, 2013). A grid search method was applied to optimize the sigma and box constraint terms as 0.3 and 1.1, respectively.

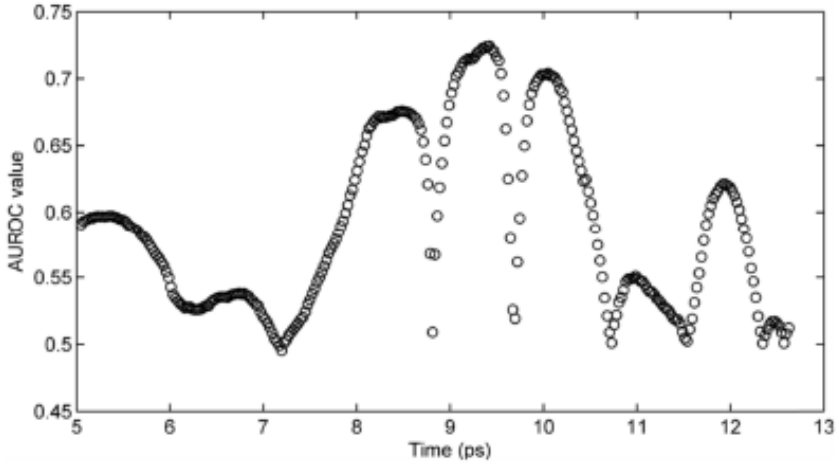


Fig. 5. AUROC analysis to evaluate the discriminative power of the amplitude parameter for time indices 5.0 – 12.6 ps. The highest AUROC values of 0.72 and 0.70 were found at t = 9.42 ps and t = 10.05 ps respectively, and these two parameters were therefore selected for tissue classification with SVM.

Tissue classification was also performed using Gaussian wavelet deconvolution in combination with a Bayesian classifier. In contrast to heuristic parameters, Gaussian wavelet deconvolution can be applied to the full impulse function. This method was considered a suitable approach because of the similarities between the signal features of a Gaussian function and its derivatives, and the TPI impulse functions from breast tissue. Gaussian derivatives of order 0 (normal Gaussian function), 1, 2, 3 and 4 were applied to the impulse function of each pixel. Higher order Gaussian derivatives were not used to avoid potential

overfitting. The Gaussian deconvolved data were then fed into a Bayesian classification algorithm [30], and classified as tumor, fibrous or adipose, respectively. Pixels classified as adipose and fibrous were then grouped together as 'benign' in order to calculate the diagnostic performance of TPI. Similar to SVM, pixels were marked as tumor when containing any amount of cancer cells.

The SVM and Bayesian classifier were trained individually using the leave-one-out method (LOO); leaving out the pixels of a single sample to be classified, and training each classifier on the other samples. The trained classifiers were then applied to the pixels of the sample that was left out. This process was repeated for all the samples, leaving each of them out in turn, and the results compiled to give accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for distinguishing malignant from benign tissue.

3. Results

3.1 Tissue sample characteristics and histopathology intra-rater reliability

In total, 126 samples from 106 patients were scanned; 46 samples from 32 patients met the strict criteria established to ensure accurate correlation of TPI with histology, i.e. a photograph was available of the sample in the histology cassette, and the number of tissue containing pixels on TPI and histopathology differed by 3 or less. These samples were included for analysis. Of these, 20 samples contained tumor; 16 invasive ductal/no special type (NST) carcinoma, 2 NST admixed with DCIS, and 2 invasive lobular carcinoma (ILC). Twenty-two samples contained pure fibrous tissue or a mixture of fibrous and adipose tissue, and 4 samples contained pure adipose tissue. The total number of pixels for analysis was 257; the breakdown in terms of tissue composition is given in Table 1.

The intra-rater reliability analysis showed excellent agreement in cell density subgroup classification between histopathological evaluation 1 and evaluation 2 ($\kappa = 0.89$) ($p = NS$). This confirmed that the established subgroups reliably reflected the tissue composition of the samples, and thus could be used to evaluate the performance of the TPI handheld probe system for different tissue groups.

3.2 Heuristic parameters and SVM classification

A total of 11 parameters were selected based on the AUROC analysis: 10 time domain parameters and 1 frequency parameter (Table 2) (Fig. 6). Most of the time domain parameters capture the area around the minimum amplitude of the pulse, and the return to baseline after the minimum. P1 – P7 were selected based on their overall ability of discriminating tumor from fibrous tissue with adipose, while P8 – P11 were specifically selected to enhance the TPI probe's ability to discriminate tumor from pure fibrous tissue. All 11 parameters showed strong discriminative power to distinguish tumor from pure adipose tissue (mean AUROC = 0.97, range 0.84 – 1.0).

Biomedical Optics EXPRESS

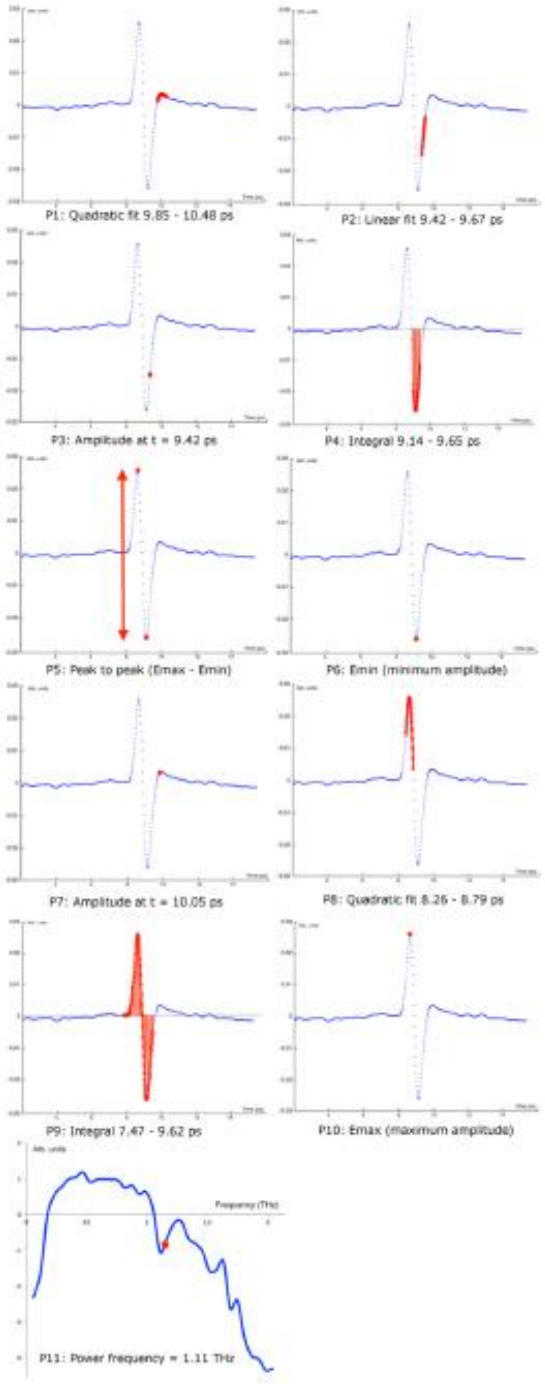


Fig. 6. Visualization of the selected parameters used in SVM classification. Each parameter is displayed in red.

Table 2. Overview of selected time domain and frequency domain parameters and their AUROC values.

Parameter	Definition	AUROC value (cell density group)*
P1	*Quadratic fit 9.85 – 10.48 ps*	*0.76 (All T and F)*
P2	*Linear fit 9.42 – 9.67 ps*	*0.73 (All T and F)*
P3	*Amplitude at t = 9.42 ps*	*0.72 (All T and F)*
P4	*Integral 9.14 – 9.65 ps*	*0.72 (All T and F)*
P5	*Peak to peak (Emax minus Emin)*	*0.71 (All T and F)*
P6	*Emin (minimum amplitude)*	*0.70 (All T and F)*
P7	*Amplitude at t = 10.05 ps*	*0.70 (All T and F)*
P8	*Quadratic fit 8.26 – 8.79 ps*	*0.74 (1 – 20% T, 100% F)*
P9	*Integral 7.47 – 9.62 ps*	*0.83 (41 – 60% T, 100% F)*
P10	*Emax (maximum amplitude)*	*0.73 (81 – 100% T, 100% F)*
P11	*Power in spectrum at frequency = 1.11 THz*	*0.82 (61 – 80% T, 100% F)*

*T = tumor, F = pure fibrous tissue or a mixture of fibrous and adipose tissue, 100% F = pure fibrous tissue only

The SVM classification results of the individual parameters, and the combination of parameters that performed best in terms of accuracy, may be found in Table 3. Overall, the combination of P1 and P6 provided the best performance with an accuracy, sensitivity, specificity, PPV and NPV of 75%, 86%, 66%, 67% and 85%, respectively. These values were obtained as a result of 16 of the 115 tumor pixels being misclassified as benign; 48 of the 142 benign pixels were misclassified as tumor. All misclassified tumor pixels had a tumor content ≤60%. Of the 48 misclassified benign pixels, 46 were fibrous pixels containing 81 – 100% fibrous cells; only 2 of the 1 – 80% fibrous pixels were misclassified as tumor, and all 26 pure adipose pixels were correctly identified as benign. The two-dimensional parametric plot of P1 and P6 showed very little differences between tumor and high percentage fibrous tissue (Fig. 7(A)); this provides an explanation for why most of the SVM classification errors occurred in these two tissue groups (Fig. 7(B)). Pixels with a high adipose content (1 – 80% fibrous pixels and pure adipose pixels) were generally clearly different from pixels containing a high percentage of fibrous tissue (81 – 100%) and cancer (Fig. 7(A)). The accuracy, sensitivity, specificity, PPV and NPV for discriminating 1 – 80% fibrous and pure adipose tissue from tumor (i.e. excluding the predominantly fibrous group with 81 – 100% purity) was 87%, 86%, 96%, 98%, 75%, respectively.

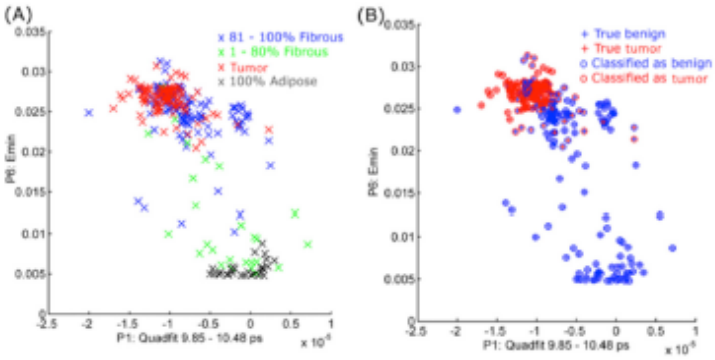


Fig. 7. Two-dimensional parametric plot (A) and SVM classification results (B) for the combination of parameters P1 and P6 that performed best in terms of accuracy.

Biomedical Optics EXPRESS

3.3 Gaussian wavelet deconvolution and Bayesian classification

The accuracy, sensitivity, specificity, PPV and NPV of Gaussian wavelet deconvolution and Bayesian classification was 69%, 87%, 54%, 60%, 84%, respectively (Table 3). Of the 115 tumor pixels, 15 were misclassified as benign. All misclassified pixels contained ≤60% tumor cells. Sixty-six of the 142 benign pixels were wrongly classified as tumor; 64 of these were 81 – 100% fibrous pixels, only two 1 – 80% fibrous pixels were misclassified. All pure adipose pixels were correctly classified. The accuracy, sensitivity, specificity, PPV and NPV of the handheld probe for discriminating 1 – 80% fibrous and pure adipose from tumor was 88%, 87%, 96%, 98% and 77%, respectively.

Table 3. Performance of heuristic parameters with SVM classification, and wavelet deconvolution with Bayesian classification. SVM classification results are shown for individual parameters and parameter combinations that performed best in terms of accuracy. The best individual parameter and parameter combination is underlined.

Parameters	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV[a] (%)	NPV[b] (%)
P1	73	87	62	65	85
P2	72	81	64	65	81
P3	70	77	65	64	78
P4	72	77	69	67	78
P5	72	92	56	63	90
P6	69	86	56	61	83
P7	69	87	54	61	84
P8	68	90	49	59	86
P9	56	69	46	51	64
P10	68	93	48	59	89
P11	56	56	57	51	61
P1 and P6	75	86	66	67	85
P1, P6 and P11	71	72	70	66	76
P1, P6, P9 and P11	67	56	75	65	68
Gaussian wavelets	69	87	54	60	84

[a]PPV = Positive predictive value

[b]NPV = Negative predictive value

4. Discussion

This study has evaluated the performance of a TPI handheld probe system to discriminate breast cancer from benign breast tissue in an *ex vivo* setting. A total of 257 pixels acquired from scanning 46 breast tissue samples were included for analysis. The tumor samples predominantly contained low-to-moderate tumor cell percentages, resembling the tissue composition found at the resection border of breast specimens from patients with positive margins after BCS. Two data analysis and classification methods were assessed: (1) heuristic parameters in combination with SVM classification and (2) Gaussian wavelet deconvolution with Bayesian classification. On the full data set the former provided the best performance in terms of accuracy (75%). Both methods had excellent sensitivity (86% and 87%, respectively) and thus show promise for identifying tumor cells close to or at the resection margins, allowing immediate further excision of appropriate margins and reducing subsequent second operations/re-excision rates if the TPI handheld probe had been used intraoperatively. Specificity however, was 66% and 54% for SVM and Bayesian respectively; for both methods the lower specificity was due to pixels with 81 – 100% fibrous tissue that were wrongly classified as tumor. The accuracy, sensitivity and specificity increased to 88%, 87%,

and 96% respectively after excluding the 81 – 100% fibrous tissue from the classification results.

The reported pooled sensitivity and specificity of the established techniques to intraoperatively assess tumor margins during BCS are 53% (95% CI 45 – 61%) and 84% (95% CI 77 – 89%) for specimen radiography, 59% (95% CI 36 – 79%) and 81% (95% CI 66 – 91%) for ultrasound imaging, 71% and 68% for radiofrequency spectroscopy, 86% (95% CI 78 – 91%) and 96% (95% CI 92 – 98%) for frozen section analysis, and 91% (95% IC 71 – 97%) and 95% (95% IC 90 – 98%) for imprint cytology, respectively [9]. Thus, based on the results of the present study, the TPI handheld probe appears to perform similarly or better in terms of sensitivity, while the performance in terms of specificity is lower. Compared to specimen radiography and ultrasound, which are also imaging technologies, TPI has the potential advantage that image interpretation is not needed as the device can provide a binary read-out (tumor or no tumor). This may overcome the need for the training required for obtaining ultrasound accreditation [31]. Potential advantages over the histopathological techniques frozen section analysis and imprint cytology are the fact that TPI is non-invasive (i.e. physical tissue disruption is not required), it does not require an on-site cytologist or histopathologist, and allows for assessing a larger tissue surface.

Results published to date on the diagnostic performance of emerging techniques for margin assessment have shown a sensitivity and specificity of 92% (95% CI 86 – 96%) and 97% (95% CI 93 – 98%) for Raman spectroscopy [10], 67 – 85% and 67 – 96% for diffuse reflectance spectroscopy [11–15], 60 – 100% and 69 – 92% for optical coherence tomography [16–18], 93 – 100% and 91.9 – 100% for mass spectroscopy [19, 20], and 87% and 76% for bioimpedance spectroscopy [21], respectively. Although the performance of the handheld probe in this study is somewhat lower than some of the other emerging techniques, TPI uses a different region of the electromagnetic spectrum and thus provides complementary information. It is possible that combinations of technologies could significantly improve the overall accuracy of identifying involved margins.

Several papers have reported on the ability of TPI to discriminate freshly excised benign from malignant breast tissue [25, 27, 32, 33]. Ashworth *et al.* performed a small pilot study using a prototype version of the TPI handheld system [32]; all other studies were conducted with systems not suited for intraoperative imaging of WLE specimens. Similar to the present results, Ashworth *et al.* found that THz impulse functions from fibrous tissue and breast cancer had strong similarities, while impulse functions from adipose tissue had clearly different features. However, none of the TPI studies in breast cancer published to date have used a data set representative of the tissue composition found at the resection border of patients with positive margins, as all tumor samples included for analysis contained >50% tumor cells. Thus, the results in our study are the first that can be used to derive insight in the potential benefits of TPI in enabling more accurate and complete tumor resection in BCS.

The accuracy, sensitivity and specificity of the TPI probe for discriminating tumor from mixed fibrous and adipose tissue, and pure adipose tissue, was 87%, 86%, and 96% for SVM, and 88%, 87% and 96% for Bayesian, respectively. Discrimination of these tissue types is most relevant clinically, as the incidence of breast cancer is higher in older women, who are likely to have fatty or mixed fibrous and fatty breasts compared to younger women who may have more dense breasts primarily composed of fibrous tissue [34].

While the results of this feasibility study are promising, two limitations were noted. Firstly, the 0.6 mm pixel distance used for correlating TPI and histopathology was based on a linear movement of the THz pulse beam across the 15 x 2 mm scan area. However, in practice the THz beam moves faster in the center of the scan window and slows down upon reaching the top and bottom boundary, resulting in a larger distance between pixels located in the center compared to the edges. This introduces a degree of inaccuracy, which was not accounted for in this study. Secondly, the current data set does not contain THz pulses from cases of pure DCIS. These samples could not be assessed, as DCIS is generally non-palpable

Biomedical Optics EXPRESS

and particularly problematic to sample in the fresh state without impairment of gold-standard histological assessment. However, since DCIS is often the cause of the clinical recommendation for re-operations in BCS, it is of key importance to assess the sensitivity of the TPI handheld probe for detecting DCIS. Based on the results of this feasibility study, a study will be performed in which TPI data will be acquired on tissue specimens with DCIS, to determine the ability of TPI to accurately detect DCIS.

In conclusion, the results of this study show that the TPI handheld probe can discriminate invasive breast cancer from benign breast tissue with a high sensitivity and an encouraging degree of accuracy. The main challenge for TPI is accurate discrimination of cancer from tissue containing a high percentage of fibrous stroma due to the similarities in the THz pulse between these two types of tissue. Larger studies are warranted to assess the performance of this technique on different tumor types including DCIS, and its impact on re-operation rate.

Compliance with ethical standards

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Funding

Guy's and St Thomas' Charity; Academy of Medical Sciences; Cancer Council Western Australia, Youngberg Women's Cancer Research Fellowship (VPW); Australian Government through the Australian Research Council (Discovery Project, DP150100635) and the National Health and Medical Research Council (Development Grant APP1074894).

Acknowledgments

The authors gratefully acknowledge the excellent support from the King's Health Partners Cancer Biobank, Breast Cancer NOW and the breast care team at Guy's Hospital for their help with patient recruitment. In particular we thank Patrycja Gazinska for digitally scanning the histopathology slides.

Disclosures

Michael Pepper and Alessia Portieri are scientific employees of Teraview Ltd; the other authors have no conflicts of interest to disclose.